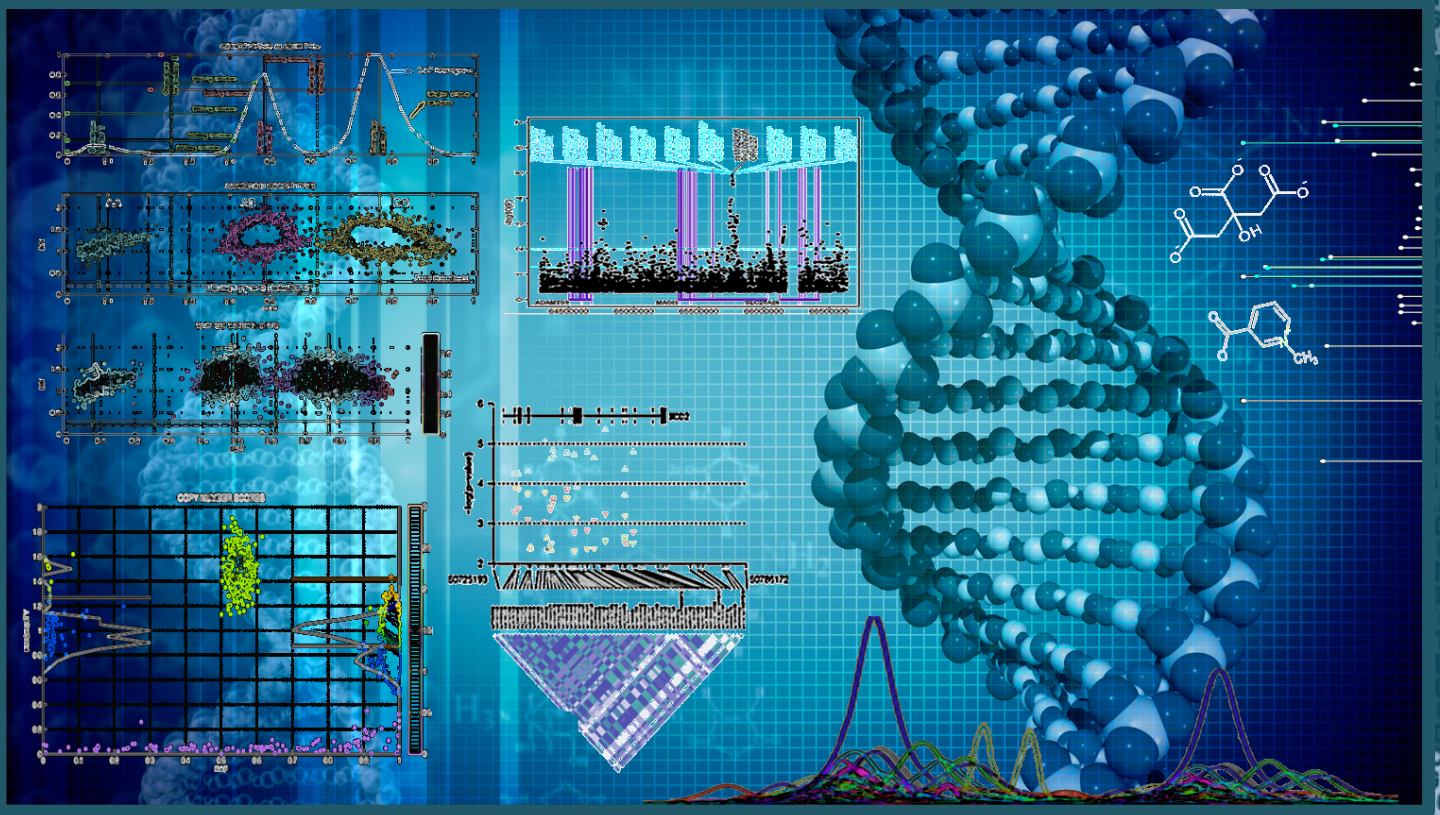
Bioinformatics methods for the genomics and metabolomics analysis of immune-mediated inflammatory diseases

Arnald Alonso Pastor



Grup de Recerca de Reumatologia

Vall d'Hebron Institut de Recerca

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola de Doctorat

Acta de qualificació de tesi doctoral

Curs acadèmic:

Nom i cognoms

Programa de doctorat
Enginyeria Biomèdica

Unitat estructural responsable del programa
Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial (ESAII)

Resolució del Tribunal

Reunit el Tribunal designat a l'efecte, el doctorand / la doctoranda exposa el tema de la seva tesi doctoral titulada

_____

_____.

Acabada la lectura i després de donar resposta a les qüestions formulades pels membres titulars del tribunal, aquest atorga la qualificació:

☐ NO APTE ☐ APROVAT ☐ NOTABLE ☐ EXCEL·LENT

(Nom, cognoms i signatura) President/a		(Nom, cognoms i signatura) Secretari/ària
(Nom, cognoms i signatura) Vocal	(Nom, cognoms i signatura) Vocal	(Nom, cognoms i signatura) Vocal

_____, _____ d'/de _____ de _____

El resultat de l'escrutini dels vots emesos pels membres titulars del tribunal, efectuat per l'Escola de Doctorat, a instància de la Comissió de Doctorat de la UPC, atorga la MENCIÓ CUM LAUDE:

☐ SÍ ☐ NO

(Nom, cognoms i signatura) President de la Comissió Permanent de l'Escola de Doctorat	(Nom, cognoms i signatura) Secretari de la Comissió Permanent de l'Escola de Doctorat

Barcelona, _____ d'/de _____ de _____

Universitat Politècnica de Catalunya
Department d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial

Vall d'Hebron Research Institute
Rheumatology Research Group

Bioinformatics methods for the genomics and metabolomics analysis of immune-mediated inflammatory diseases

Dissertation submitted for the degree of
Doctor of Philosophy in Biomedical Engineering

Author:
Arnald ALONSO PASTOR

Thesis supervisors:
Antonio JULIÀ CANO
Sara MARSAL BARRIL
Thesis tutor:
Alexandre PERERA LLUNA

2015 Barcelona

Arnald ALONSO PASTOR
Rheumatology Research Group
Vall d'Hebron Institute of Research

This thesis was supported by:

PhD fellowship AGAUR-FI-2013/00974

Generalitat de Catalunya

It is not his possession of knowledge, of irrefutable truth, that makes the man of science, but his persistent and recklessly critical quest for truth.

Karl R. Popper

El hombre es el animal que pregunta. El día en que verdaderamente sepamos preguntar, habrá diálogo. Por ahora las preguntas nos alejan vertiginosamente de las respuestas.

Julio Cortázar

Abstract

During the last decade, genomics have been widely used to the characterization of the molecular basis of common diseases. Genome-wide association studies (GWAS) have been highly successful in characterizing the genetic variation that influences human traits including the susceptibility to common diseases. In metabolomics, recent improvements of analytical technologies have enabled the analysis of complete metabolomic profiles. Using this approach, high-throughput metabolomics studies have already demonstrated a high potential for the discovery of disease biomarkers.

The use of powerful high-throughput measurement technologies has resulted in the generation of large datasets of biological variation. In order to extract relevant information from this data, highly specialized bioinformatics methods are required. This thesis is focused on the development of new methodological tools to improve the processing of genomics and metabolomics high-throughput data. These new tools have been used in the analysis framework of the Immune-Mediated Inflammatory Diseases (IMIDs) Consortium. The IMID Consortium is a large Spanish network of biomedical researchers on autoimmune diseases, which holds one of the largest collections of biological samples from this group of diseases, as well as healthy controls.

The first analysis tool that has been developed is a computationally efficient algorithm for the simultaneous genotyping of single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) using microarray data. This bioinformatics tool, called GStream, integrates the genotyping of both types of genomic variants into a single processing pipeline. We demonstrate that the developed algorithms provide a significant increase in genotyping accuracy and call rate when compared to previous algorithms. Using GStream, researchers performing large-scale GWASs will not only benefit from the combined and fast genotyping of SNPs and CNVs but, more importantly, they will also improve the accuracy and therefore the statistical power of their studies.

The second tool that was developed during this thesis was FOCUS, a bioinformatics

framework that provides a complete data analysis workflow for high-throughput metabolomics studies based on one-dimensional nuclear magnetic resonance (NMR). FOCUS workflow includes quality control, peak alignment, peak picking and metabolite identification. The algorithms included in FOCUS were designed to overcome several technical challenges of NMR that can dramatically affect the quality of the results. FOCUS allows users to easily obtain high-quality NMR peak feature matrices, which are ready for chemometric analysis, as well as metabolite identification scores for each peak that greatly simplify the biological interpretation of the results. When tested against previous NMR data processing methodologies, FOCUS clearly showed a superior performance, even in datasets with high levels of spectral unalignment.

The final research work included in this thesis is a GWAS in Crohn's disease (CD) clinical phenotypes. CD is the most prevalent chronic inflammatory disease of the bowel, and is characterized by segmental and transmural inflammation of the gastrointestinal tract. CD is a highly heterogeneous disease, with patients showing different degrees of severity. The identification of the genetic basis associated with disease severity is therefore a major objective in CD translational research. The present PhD thesis includes the first GWAS of clinically relevant phenotypes in CD. A total of 17 phenotypes associated with different clinical complications were analyzed. In this study, we identified new genetic regions significantly associated to complicated disease course, disease location, mild disease course, and erythema nodosum. These findings are of high relevance since they show, for the first time, the existence of a genetic component for disease heterogeneity that is independent of the genetic variation associated with susceptibility to CD.

Resum

Durant la darrera dècada, la genòmica ha jugat un paper clau en la caracterització de la base molecular de les malalties complexes. Els estudis d'associació de genoma complet (GWAS) han permès la caracterització de les regions genètiques que influencien fenotips humans tals com la susceptibilitat a desenvolupar una malaltia complexa. En metabolòmica, millores recents en les tecnologies analítiques han impulsat l'obtenció de perfils metabolòmics en grans cohorts de mostres. Els estudis resultants han demostrat també un gran potencial per al descobriment de biomarcadors d'utilitat en malalties humanes.

L'aplicació de les tecnologies *high-throughput* permet generar grans conjunts de dades de variació biològica. Per tal d'extreure la informació de rellevància d'aquestes dades es requereix l'aplicació de potents eines bioinformàtiques. Aquesta tesi està centrada en el desenvolupament de nous mètodes per a millorar i agilitzar el processament de dades genòmiques i metabolòmiques *high-throughput*, així com la seva posterior implementació en forma d'aplicacions bioinformàtiques. Aquestes noves eines han estat incorporades al flux d'anàlisi del consorci IMID (malalties inflamatòries mediades per immunitat). Aquest consorci és una xarxa espanyola que aglutina un gran nombre d'investigadors biomèdics amb l'interès comú de l'estudi de malalties autoimmunes. El consorci IMID disposa d'una de les col·leccions més extenses de mostres biològiques de pacients d'aquest grup de malalties, així com d'individus sans control.

La primera eina bioinformàtica implementada consisteix en un conjunt d'algoritmes que integren el genotipat de polimorfismes de nucleòtid simple (SNPs) i variacions de nombre de còpies (CNVs) sobre dades de microarrays de genotipat. Aquesta eina, anomenada GStream, incorpora de forma eficient tot el flux d'anàlisi necessari per al genotipat en GWAS. S'ha demostrat que els algoritmes desenvolupats milloren significativament la precisió del genotipat i augmenten el nombre de variants genètiques identificades respecte a les metodologies anteriors. La utilització d'aquesta eina permet doncs ampliar el nombre de variants genètiques analitzades i correctament genotipades, incrementant de forma significativa el poder estadístic dels estudis genètics GWAS.

La segona eina desenvolupada durant el curs d'aquesta tesi ha estat FOCUS. Es tracta d'una eina bioinformàtica integrada que inclou totes les etapes de processat d'espectres de ressonància magnètica nuclear (NMR) per a estudis de metabolòmica. El flux d'anàlisi inclou el control de qualitat, l'alineament/quantificació de pics espectrals i, finalment, la identificació dels metabolits associats a cadascun dels pics quantificats. Tots els algoritmes han estat dissenyats per a corregir els biaixos i errors que sovint limiten considerablement la qualitat dels resultats i que són un dels reptes tècnics no resolts més importants de la metabolòmica actual. FOCUS obté directament una matriu numèrica d'alta qualitat llesta per a l'anàlisi quimiomètric, i genera uns *scores* d'identificació que simplifiquen la interpretació biològica dels resultats. FOCUS ha assolit un rendiment significativament superior al de metodologies prèvies.

Aquesta tesi conclou amb el primer GWAS de fenotips clínics de malaltia de Crohn. Aquesta malaltia IMID és la malaltia inflamatòria intestinal de major prevalença, i es caracteritza per la inflamació del tracte gastrointestinal. La malaltia de Crohn és molt heterogènia, amb pacients que presenten graus molt diferents de gravetat. La identificació de variants genètiques associades als fenotips d'aquesta malaltia és, per tant, un dels objectius més rellevants per a la investigació translacional. Un total de 17 fenotips han estat analitzats utilitzant cohorts de descobriment i validació per tal d'identificar i replicar loci de risc associats a cadascun d'ells. Els resultats de l'estudi han permès identificar, per primer cop, regions genètiques associades a l'evolució de la malaltia, la localització de l'afectació i les manifestacions extraintestinals. Aquests resultats són de gran rellevància ja que no tan sols han permès identificar noves vies biològiques associades a diferents fenotips clínics, sinó que també demostren, per primer cop, la existència d'un component genètic de la heterogeneïtat a la malaltia de Crohn i que és independent de la variació genètica associada al risc de patir la malaltia.

Acknowledgements

La tesi doctoral que es presenta a continuació ha estat un treball complex i llarg, un camí ple d'alegries i entrebancs en el què cada problema ha suposat un gran repte i en el què cada solució ha estat possible gràcies a tota la gent que m'ha envoltat, m'ha aconsellat, s'ha involucrat i ha contribuït i confiat en la feina que s'estava fent. Al llarg d'aquest camí, ja sigui des d'una perspectiva més professional, personal o ambdues alhora, el suport continu o puntual de totes aquestes persones ha estat clau en la superació de tots els reptes que han sorgit, clars partícips també de les alegries per cada objectiu assolit i de la satisfacció de poder concloure aquesta etapa amb la sensació d'una bona feina feta.

Primer de tot vull agrair als meus directors de tesi, el Dr Toni Julià i la Dra Sara Marsal, el suport, la confiança, el guiatge i la formació que m'han ofert durant tots aquests anys. Són ells sens dubte els que m'han encomanat aquesta passió per la recerca necessària per arribar amb èxit al punt en que sóc ara. La seva contribució en tots els treballs realitzats ha estat sempre activa, constructiva i fonamental en tots els àmbits d'aquesta tesi doctoral. D'ella n'he tret un aprenentatge continu i molt valuós, tant de coneixements com del propi procés que porta a assolir un objectiu específic de recerca científica. També voldria donar les gràcies al meu tutor, el Dr Alexandre Perera, pels consells i el guiatge que m'ha ofert durant la primera etapa del Màster d'Enginyeria Biomèdica i durant la realització de la tesi doctoral.

Tota aquesta feina també ha estat possible gràcies a tot l'equip que forma el Grup de Recerca de Reumatologia. Gràcies a la feina de tots plegats, al dia a dia, s'han superat amb èxit tots els projectes de recerca en els que m'he vist involucrat. Així, personalment, voldria agrair a la Maria, el Raül, la Núria, l'Elena, l'Adrià, el Pablo i la Carla la seva feina i la seva disposició sempre oberta a resoldre qualsevol dubte o donar un cop de mà. Aquí també voldria agrair a la Gaby i l'Isa l'aprenentatge dels treballs conjunts, la seva actitud professional i personal ha estat sempre un bon revulsiu aquests anys.

A totes les persones que formen el consorci IMID, sense la dedicació dels quals aquests projectes de recerca biomèdica a nivell nacional no serien possibles. Personalment m'agradaria agrair al Dr Eugeni Domènech i al Dr Javier P. Gisbert la seva dedicació i participació activa i necessària per al desenvolupament del GWAS de fenotips de Crohn.

A tot el grup de la plataforma de metabolòmica de la Universitat Rovira i Virgili. Especialment al Dr Xavier Correig i la resta de membres de l'equip, com la Mariona i el Miguel, per la seva col.laboració imprescindible en els estudis de metabolòmica que hem realitzat durant aquests anys.

Per estar a l'altre cantó d'aquesta tesi doctoral, família i amics. Als pares, per tots els esforços que han realitzat per a que pugui arribar on sóc ara, per la seva fortalesa, la seva paciència i el seu suport continu, per transmetre'm sempre les forces necessàries, per Vallclara, per París i per estar-hi sempre. Als meus germans, l'Edu i l'Anna, als tiets, tietes i cosins. Molt especialment també al meu avi. A tots els amics i amigues amb els que he compartit tants moments, per les llargues tardes a les terrasses de Sant Antoni, Osca i Collblanc. A tota la colla del màster de biomèdica, de telecos i París. A la meva tot terreny amb pedals, pels moments de pau tot travessant les muntanyes de Prades, la serra del Montsant i les planes de la Conca de Barberà.

Finalment només em resta agrair a la que ha estat la persona més important durant aquests anys, la Cristina. Per transmetre'm sempre la seva vitalitat, per escoltar-me, per la seva paciència i el seu suport incondicional, pels somriures, l'afecte i els moments viscuts, d'on he tret les forces necessàries als moments més difícils. Moltes gràcies.

Contents

List of Figures

List of Tables

Abbreviations

1KGP 1000 Genomes Project. 13, 15, 39, 40, 49, 52, 55–57, 64

BAF B-Allele frequency. 42, 44, 46, 50

bp Basepair. 5, 10–12, 20

CD Crohn's disease. vi, xiii, xvi, 1–3, 35, 36, 38, 88–95, 97, 100–103, 105, 108, 109, 111

CDC Complicated disease course. 89, 93, 95, 96, 102

CGH Comparative genomic hybridization. 5, 39–41, 50, 51, 64, 65

CNP Copy Number Polymorphism. xvi, xix, 40, 50, 51, 58, 59

CNV Copy Number Variant. v, vii, xi, xii, xv, xix, 2, 5, 11, 12, 16, 21–23, 36, 37, 39–47, 49–53, 55–58, 60–66, 105, 106, 111

DNA Deoxyribonucleic acid. xv, 5–13, 15, 16, 22, 35

GC Gas chromatography. 26, 29, 30

GMM Gaussian Mixture Model. 45, 47

GRR Grup de Recerca de Reumatologia (Rheumatology Research Group). 1

GWAS Genome-Wide Association Study. v–viii, xiii, xix, 1–3, 5, 15, 17–20, 35, 36, 38–41, 52, 54, 60, 61, 63–66, 87–93, 95, 96, 100, 101, 103, 105, 106, 108, 109, 111

HGP Human Genome Project. 5

HMM Hidden Markov model. 22, 50

IBD Inflammatory Bowel Disease. 1

IMID Immuno-mediated inflammatory disease. v, vii, viii, 1, 3, 35, 112

IMIDC IMID Consortium. 1

LC Liquid chromatography. xv, 26, 29–31, 33

LD Linkage Disequilibrium. 12–14, 41, 106

m/z Mass-to-charge ratio. 26, 29, 31, 32

MAF Minor allele frequency. 10, 11, 13, 15, 21, 42, 48, 49, 54, 95, 97, 98

mRNA Messenger RNA. 7, 8

MS Mass spectrometry. xv, 24, 26, 29–33

NCBI National Center for Biotechnology Information. xv, 10, 11, 47

NMR Nuclear magnetic resonance. v, vi, viii, xv, 2, 24–33, 37, 67–71, 73, 77, 78, 81–83, 85, 86, 107, 108, 112

PDF Probability density function. 42–44

ppm Parts per million. 24, 26, 29, 30, 32

PS Psoriasis. 1

PsA Psoriatic arthritis. 1

QC Quality control. 33, 91, 92

RA Rheumatoid arthritis. 1

RNA Ribonucleic acid. xv, 6–8

SLE Systemic Lupus Erythematosus. 1

SNP Single Nucleotide Polymorphism. v, vii, xi, xii, xv, xix, 2, 5, 9–11, 13–22, 36, 37, 39–48, 50, 52–55, 60, 61, 63–66, 105, 106, 109, 111

UC Ulcerative colitis. 1

VHIR Vall d'Hebron Institute of Research. 1

1 | Introduction

1.1 Thesis outline

The current thesis has been developed by the student Arnald Alonso at the Rheumatology Research Group of the Vall d'Hebron Hospital Institute of Research (GRR-VHIR), under the supervision of Dr. Sara Marsal and Dr. Antonio Julià.

This thesis has taken place during the execution of the singular and strategic project *IMID-Kit* (PSE-010000-2006-6, Ministerio de Economía y Competitividad). This project, coordinated by Dr. Sara Marsal (GRR-VHIR), is a multicentric collaborative project of the IMID Consortium (IMIDC). The IMIDC is a Spanish network of researchers that includes more than 80 clinical departments of university hospitals from Spain, and is focused in the study of immune-mediated inflammatory diseases (IMIDs) through the analysis of high-throughput molecular (omics) data. IMIDs are a group of prevalent autoimmune diseases of unknown etiology that are characterized by sharing common inflammatory pathways (i.e. TNF-α pathway[1]). These chronic diseases are known to produce a relevant socio-economic burden due to their increased use of healthcare resources, as well as the reduction of quality of life and productivity of IMID patients[2]. One of the main milestones of the *IMID-Kit* project has been the collection of biological samples and associated clinical information from more than 13,000 patients from six of the most prevalent IMIDs: rheumatoid arthritis (RA), psoriasis (PS), psoriatic arthritis (PsA), Crohn's disease (CD), ulcerative colitis (UC) and systemic lupus erythematosus (SLE). The analysis of the genomic and metabolomic data obtained from these large patient cohorts has been the starting point for the development of this thesis.

This thesis is divided in two sections. The first section is based on the development of new algorithms and software tools for the processing of genomic and metabolomic high-throughput data. The main objectives of these methods were to improve the accuracy and usability. In the second section, the first genome-wide association analysis (GWAS) of CD clinical phenotypes, the most prevalent inflammatory bowel disease (IBD), has been performed.

In the first section, an algorithm (GStream, *Chapter 3*) for the simultaneous genome-wide genotyping of single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) has been developed. This algorithm has been designed to solve different technical challenges associated with SNP and CNV genotyping in GWASs based on genotyping microarrays: (1) the absence of bioinformatics tools that integrate the simultaneous genotyping of both types of polymorphisms, and (2), the low performance of previous CNV genotyping methods in microarray data. GStream efficiently integrates the whole genotyping process (quality control, SNP and CNV genotyping) in a single software tool. The genotyping performance has been quantitatively compared against previous state-of-the-art methods using different Illumina genotyping microarrays, together with publicly available SNP and CNV reference datasets.

Second, a bioinformatics tool for the processing of metabolomics data (FOCUS, *Chapter 4*) has been implemented to improve the quantification of spectral peaks in metabolomics data from nuclear magnetic resonance (NMR) studies. FOCUS software was designed to solve several technical challenges that critically limit the application of NMR to high-throughput metabolomics studies. These limitations are the introduction of significant biases in peak positions associated with the sample chemical environment, and the lack of reliable automatic metabolite identification methods. The most important contributions of FOCUS have been: (1) the inclusion of a new spectral alignment algorithm that significantly outperforms previous methods, (2) a new metabolite identification algorithm that uses multiple peak features to perform metabolite assignment, and (3), the integration of the entire NMR data processing workflow in a single user-friendly software tool.

The two bioinformatics tools implemented in this first section of the thesis -GStream and FOCUS- have been published in top-tier journals (*Plos One*[3], IF: 3.53; *Analytical Chemistry*[4], IF: 5.83) and are distributed as open-source software tools (*www.urr.cat/software*).

The second section of this thesis is devoted to the analysis of the omics data from patients included in the *IMID-Kit* project. In this study the work performed by the PhD student has allowed the identification of several genetic variants associated to Crohn's disease (OMIM 266600) phenotypes (*Chapter 5*). CD is a prevalent (20-150 cases per 100,000 persons[5]) chronic inflammatory disease characterized by the chronic inflammation of the gastrointestinal tract[6]. This persistent intestinal inflammatory activity leads to multiple disease-related complications like intestinal stenosis, fistulas, and abscesses that significantly reduce the quality of life of CD patients[6;7]. From a clinical perspective, the study of the more severe disease phenotypes, such as those that require bowel resection, is of high importance. These clinically relevant phenotypes have shown a significant level of aggregation within individual families[8;9], thereby suggesting that there is a genetic basis for disease heterogeneity. To

date, several genetic association studies[10–14] have investigated the association of previously known CD risk loci with disease phenotypes, but only genetic variants in *NOD2* locus have been consistently associated[15;16]. The present work represents the first GWAS of clinically relevant phenotypes in CD, which include disease location, disease behavior, disease course, age at onset and extraintestinal manifestations.

The results of this study have been published in the top-tier journal of the medical specialty (*Gastroenterology*[17], IF: 13.93).

In addition to the previous works, the PhD student Arnald Alonso has also been actively involved in the elaboration of two review articles in metabolomics. The first manuscript provides a state-of-the-art in metabolomics methodology (*Alonso et al. 2015, Appendix D.1*), and the second manuscript comprehensively reviews the most recent results of this high-throughput technology in the study of IMIDs (*Julià et al. 2014, Appendix D.2*).

Figure 1.1 shows the distribution over time of the different projects and studies related with this thesis.

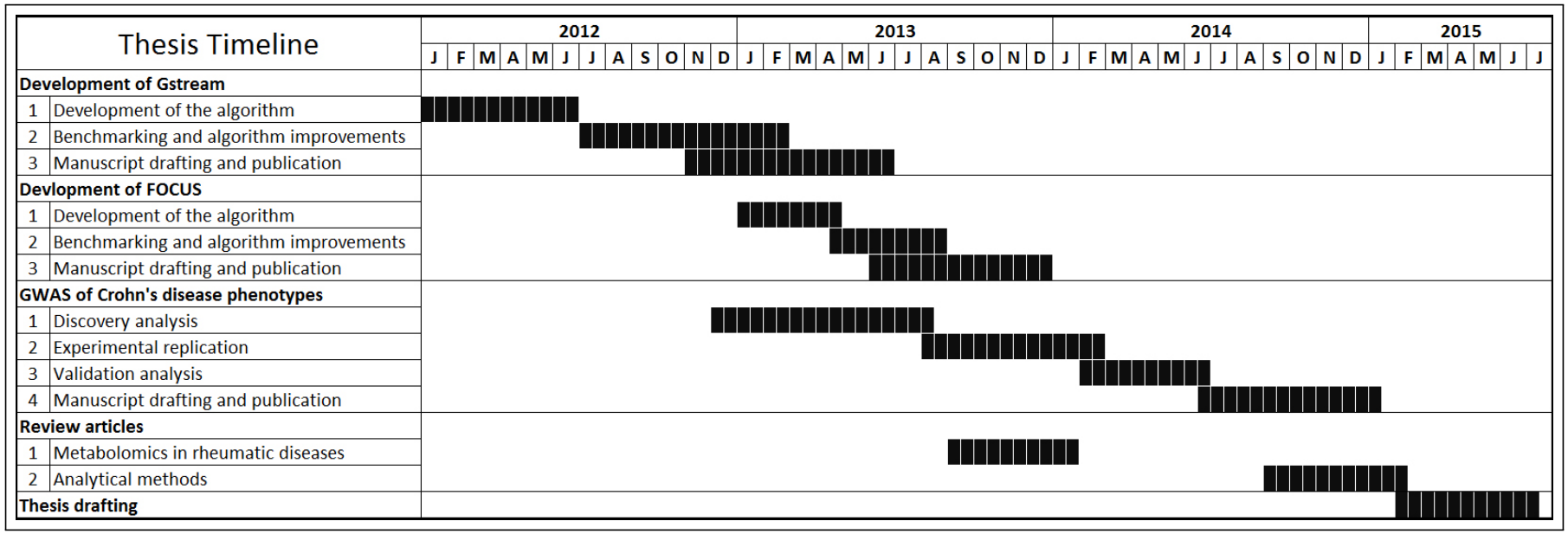


Figure 1.1: *Thesis timeline.* The current PhD Thesis was started in January of 2012 and was completed in July of 2015. During this period, the PhD student has been involved in several projects that are shown in this timeline.

1.2 Genomics

Genome-wide association studies (GWAS) have played an important role in the identification of associations between common genetic variants and the susceptibility to complex diseases[18–21]. The development of the microarray technology in the early 2000s was a key point for the success of GWASs. Genotyping microarrays[22] are technological platforms designed to capture most of the common variation of the human genome by simultaneously genotyping thousands of SNPs[23]. In addition to SNPs, CNVs are also an important source of genetic variation[24]. CNVs are defined as DNA segments larger than 1,000 basepairs (bps) that are present at variable copy number in comparison with a reference genome[25]. Recent studies have demonstrated their importance based on their high frequency, functional impact and role in human disease[26–30]. The development of genotyping microarray and comparative genomic hybridization (CGH) technologies has enabled the inclusion of CNVs in GWAS analyses[31]. Compared with CGH, genotyping microarrays have the advantage of allowing the genotyping of both SNPs and CNVs. Nevertheless, since genotyping microarrays were initially designed for SNP genotyping, powerful algorithms were needed to deal with the technical challenges of CNV genotyping. The absence of such methodological tools motivated the development of GStream genotyping software (*Chapter 3*).

The following subsections provide an introduction to the genomics concepts that are directly related with the work developed during this thesis.

1.2.1 The human genome

The human genome consists of approximately $3 \cdot 10^9$ DNA bps distributed in 23 chromosome pairs that are located in the cellular nucleus. From these, 22 are autosomal chromosomes (i.e. chromosomes 1 to 22), and the remaining two chromosomes are the sex chromosome pair (XX in women and XY in men). In addition to these 23 chromosome pairs, the human genetic information consists also of mitochondrial chromosomes that are located in the mitochondria. Mitochondrial chromosomes are circular ($\sim$16,600 bps), and are only inherited from the mother.

The Human Genome Project (HGP)[32] was crucial for the characterization of the human genome sequence. This project took place between 1990 and 2003, and involved a large effort of multiple public research centres to complete the human DNA sequence[33]. In parallel, a similar, privately funded project was also performed by the company Celera Genomics, created by the biology researcher Craig Venter[34]. Once merged, both datasets constituted the first reference sequence of the human genome, clearly one of the major achievements of science.

The DNA sequence of an individual provides its genetic information and influences the regulation and development of the organism. The DNA sequence contains the information required to synthesize ribonucleic acid (RNA) and protein molecules, which perform and regulate all the cellular functions that, at a higher organizational level, provide functionality to the tissues and organs of an individual. DNA molecules are composed of two nucleotide strands that form a double helix (Figure 1.2). Nucleotides are the base units of DNA, and are made up of a sugar, a phosphate group and one of the four following nitrogenous bases: adenine (A), thymine (T), guanine (G) or cytosine (C). Nitrogenous bases characterize the genomic information that provides each nucleotide which, in turn, characterizes the DNA sequence of an organism. Both DNA strands are bound together by hydrogen bonds, which occur between each base on one strand and its complementary on the other strand. Only two combinations of nitrogenous bases can be bound together by hydrogen bonds: adenine-thymine (2 H bonds) and cytosine-guanine (3 H bonds). Therefore, the nucleotide sequence of one strand automatically determines the sequence of the complementary strand.

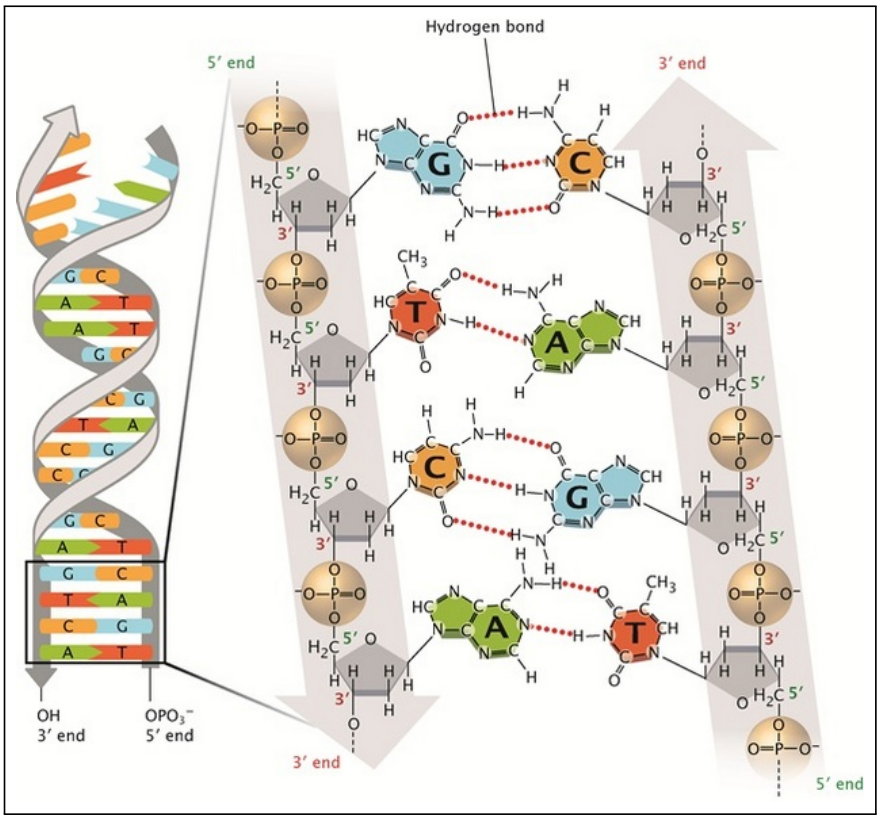


Figure 1.2: *DNA molecule.* DNA structuctural units: sugar (S), phosphate group (P) and nitrogenous base. Nucleotides of the same strand bind together by phosphodiester bonds while both strands bind together by hydrogen bonds. As shown in the figure, hydrogen bonds determine the specificity of base pairing: adenine-thymine (2 H bonds) or cytosine-guanine (3 H bonds) nucleotides. Source: *Pray L et al*[35].

Genes are defined as segments of the genome that are copied into ribonucleic acid (RNA) molecules. This biochemical process is performed by RNA polymerase enzymes and is

known as transcription (Figure 1.3). Messenger RNA molecules (mRNAs) are the group of RNAs that provides the information to synthesize proteins through the process called translation (Figure 1.3). mRNA translation occurs in the ribosome, a complex molecular structure made of (ribosomal)RNA and proteins. Ribosomes build up proteins by translating groups of three consecutive nucleotides (i.e. codons) into specific amino acids, a unique relation known as the genetic code. Together, transcription and translation are known as the central dogma of biology, and its discovery lead to the still ongoing biotechnological revolution.

To date, more than 20,000 protein-coding genes have been described in the human genome[36;37]. The structural units that compose genes are (Figure 1.3):

- Exons: The nucleotide sequences of the gene that will be translated into an amino-acid sequence by ribosomes.
- Introns: DNA sequences that are copied into mRNA molecules but are not translated into amino-acids. These regions are placed between exons and are removed from the RNA molecule through a biochemical cut-and-paste process known as splicing.

Exons only comprise 2.8% of the human genomic DNA. The remaining genomic sequence is categorized as intronic (42.3%) or intergenic (54.9%)[39]. Although intronic and intergenic regions do not code for proteins, they have increasingly shown to be crucial for the regulation of genetic activity in cells:

- Non-coding RNA: RNA segments that are transcribed but not translated. This type of RNA includes long and small non coding RNAs. Non-coding RNAs have a wide range of functions such as gene expression regulation (Figure 1.4)[40].
- Regulatory elements: DNA sequences that are the binding sites of proteins that control the rate of transcription of genetic information from DNA to mRNA (i.e. transcription factors)[39]. The most abundant regulatory elements in the genome are promoters, enhancers and silencers. Promoters are sequences immediately adjacent to the beginning of the gene (*cis*-regulatory elements), while enhancers and silencers are gene expression regulation sequences that can act on distant genomic locations or even in different chromosomes (*trans*-regulatory elements).

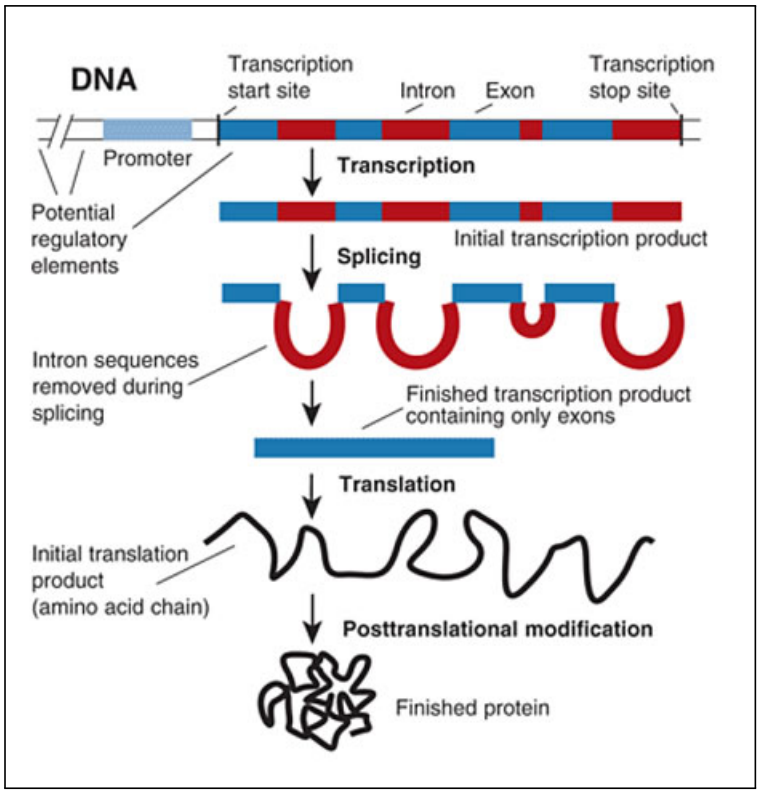


Figure 1.3: *Structural organization of the genome and transcription/translation processes.* In transcription, the DNA sequence is copied to the mRNA molecule. Non coding regions (introns) are further removed, and exons are then assembled (splicing). The resulting mRNA molecule is then translated into an amino-acid sequence that will lead to an active protein. Source: *Hiller S et al*[38].

1.2.2 Human genome variation

The DNA sequence of the human genome is different between individuals. These differences in sequence are known as genomic or genetic variants, and are major contributors to the phenotipic variability observed in humans, including the risk to develop a certain disease.

From a classical genetics point of view, human diseases can be grouped into two groups: mendelian diseases and complex diseases. Mendelian diseases are diseases caused by a single, highly penetrant mutation, and are characterized by clear inheritance patterns within families. Complex diseases, instead, are caused by the aggregated effect of multiple genetic variants and their interaction with different environmental factors (Figure 1.5).

Single nucleotide polymorphisms

A SNP is a single DNA basepair variation, where two or more different nucleotides can be observed within a given population. Like any other DNA polymorphism, the set of observed nucleotides in a SNP are called alleles. Figure 1.6 shows an example of this type of variation where the two observed SNP alleles correspond to a cytosine (C) and a guanine (G). In most

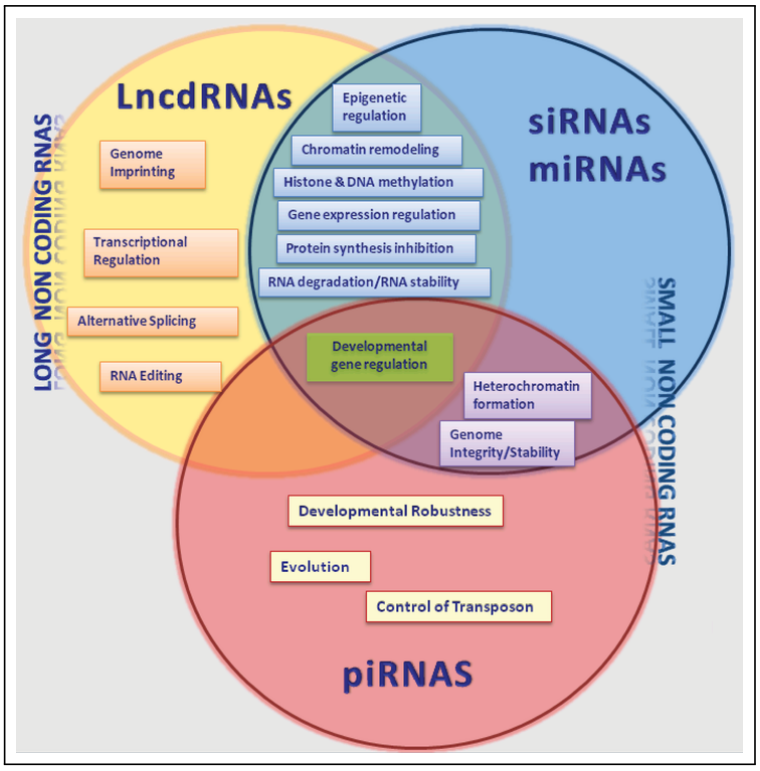


Figure 1.4: *Non-coding RNAs.* Types of non-coding RNAs. Source: *Quintal A et al*[40].

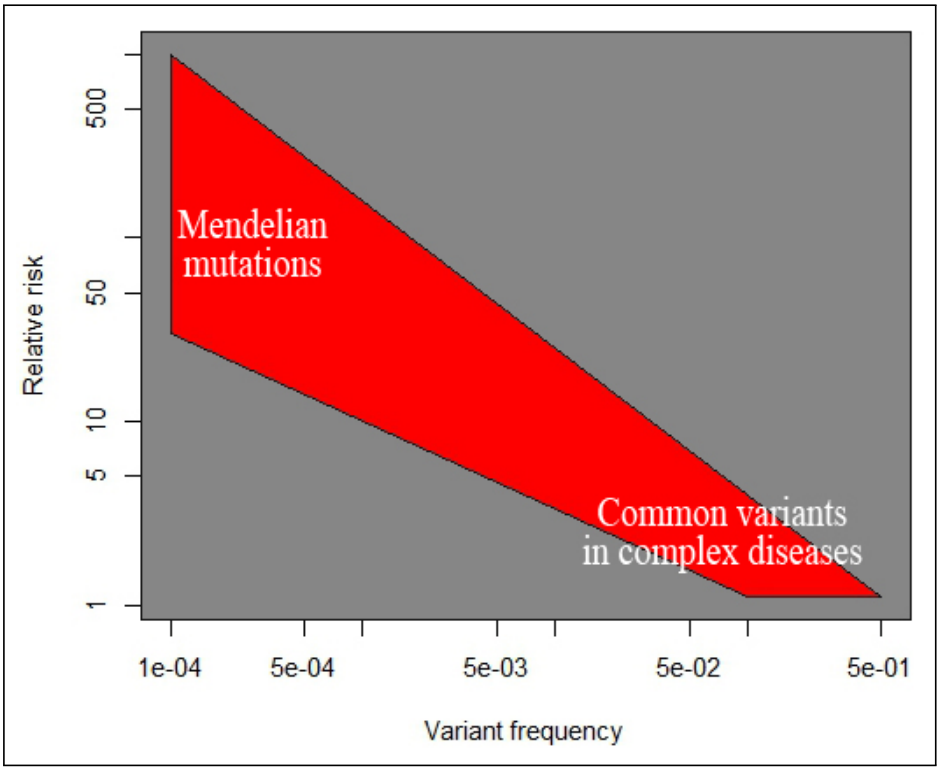


Figure 1.5: *Spectrum of genetic mutations and variations.* Mendelian mutations are rare and highly penetrant for an inherited disorder. Conversely, common genetic variants confer a much smaller risk for the phenotype, and interact with other genetic variants and environmental factors to produce complex diseases.

SNPs only two alleles are identified and therefore, they are generally considered biallelic. Given that the human genome is diploid (i.e. there are two copies of each chromosome), three possible allele combinations (genotypes) can be observed in a single individual. For the previous case, the three possible genotypes are two homozygotes (CC and GG individuals) and one heterozygote (CG or GC individuals).

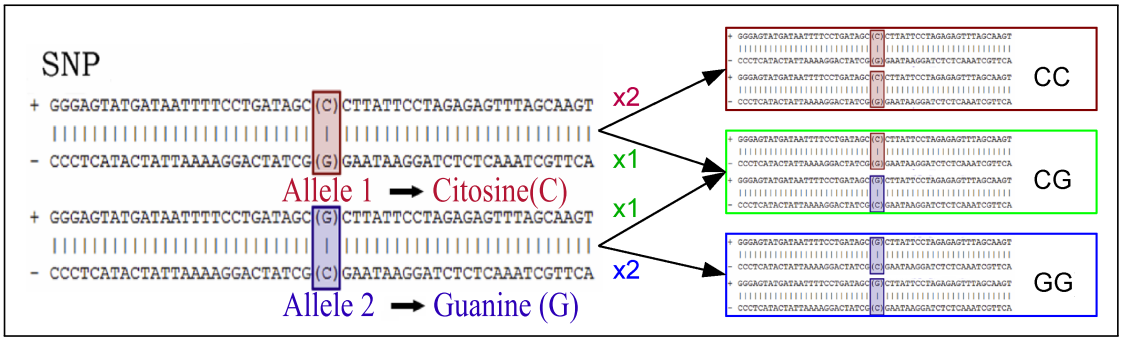


Figure 1.6: *Single nucleotide polymorphism.* The left panel shows the two sequence variants (alleles) at a single DNA basepair. The right panel shows how these two alleles lead to three genotype possibilities in diploid organisms. The alleles are generally called according to the positive DNA strand.

A SNP is defined by its physical position in the chromosome (i.e. basepair) and by its alleles. Depending on the frequency of each allele in a given population, SNP alleles are referred as the minor or major alleles. The information of the allele frequencies of one SNP is usually reported using only the minor allele frequency (MAF). SNPs are relevant for the study of the genetic basis of human diseases since they are highly abundant in the human genome (approximately 1 SNP every 100-200 bps). The reference SNP database (dbSNP, *www.ncbi.nlm.nih.gov/snp*) maintained by the National Center for Biotechnology Information (NCBI, USA) stores information about a total of 116 million human SNPs (dbSNP Build 142; Figure 1.7). From these, 47.6 million SNPs (41.0%) are located in coding regions, 45.4 million in introns (39.1%) and 2.2 million in exons (1.9%). According to their functional impact, exon SNPs are classified into three main categories:

- Synonymous SNPs if they do not lead to a change in the amino-acid sequence of the resulting protein.
- Non-synonymous missense SNPs if they lead to an amino-acid change in the resulting protein.
- Non-synonymous nonsense SNPs if they lead to a gain or a loss of a stop codon that produces a shortened or elongated version of the protein, respectively.

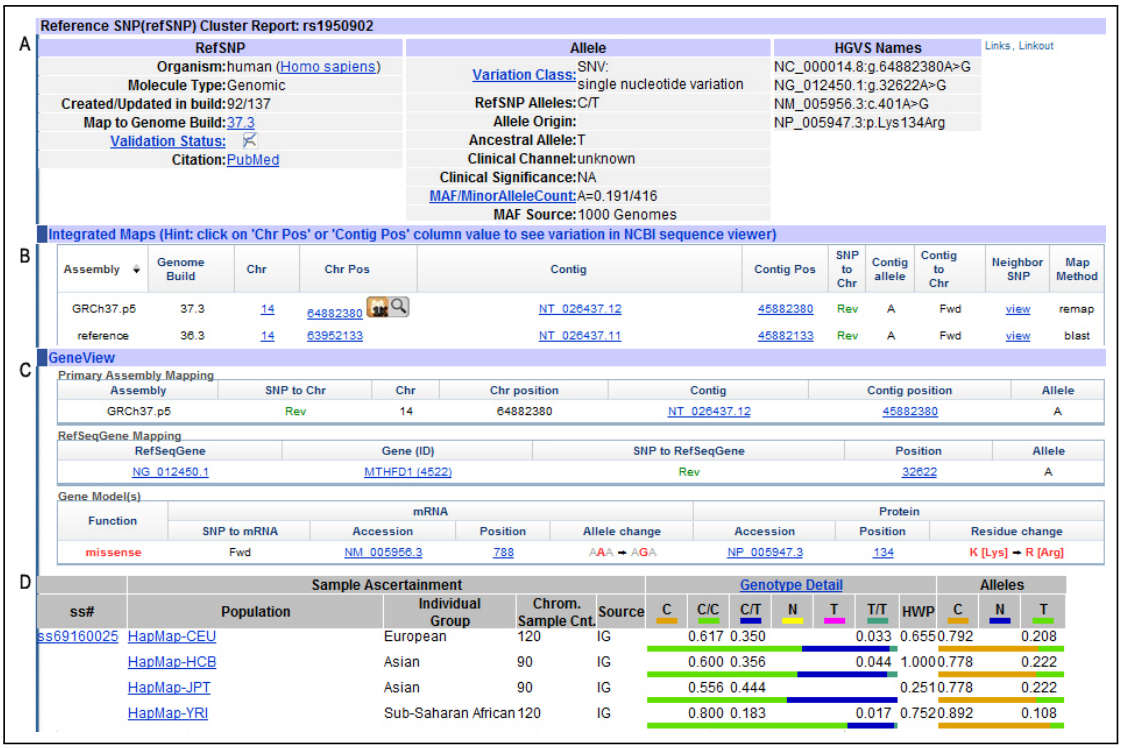


Figure 1.7: *SNP properties as described in NCBI dbSNP database.* This figure shows an example of the data that characterizes one particular SNP (i.e. SNP rs1950902). The upper panel (A) provides information about the SNP alleles and their corresponding MAFs. The second panel (B) gives the SNP position within the genome and the third panel (C) shows the genomic context of the SNP. In this example, the third panel shows how the SNP is located in the coding region of *MTHFD1* gene. Finally, the fourth panel (D) gives information about the allelic and genotypic frequencies of the SNP in different reference populations.

Structural variants

In addition to SNPs, structural variants are also a common source of genetic variability in the human genome[41]. When multiple human genomes are compared, the number of basepair differences due to structural variation are higher than those due to single basepair mutations[26–30;42]. Structural variants are ubiquitous in the human genome, and have been associated to the development of Mendelian as well as complex diseases[43;44].

Structural variants can be classified into three categories (Figure 1.8):

- Copy number variants (CNVs): This type of variation is characterized by gains or losses of long DNA sequences (from 100 bps up to thousands of bps). There exist three CNV types:
  - Deletions: Loss of a DNA segment with regard to the reference genome.
  - Amplifications: Gain of one or more copies of a DNA segment with regard to the reference genome.

  - Insertions: A new DNA segment that is not present in the reference genome is added to the DNA sequence.

- Tandem repeats: DNA segments characterized by the adjacent repetition of two or more nucleotides. Tandem repeats can be classified into microsatellites (repeated motif < 10 bps) or minisatellites (repeated motif > 10 bps).

- InDel: Similar to CNVs, they refer to gains or losses of DNA segments, but affecting a much shorter sequence of DNA (< 100 bps).

- Inversions: Chromosomal rearrangements where a DNA sequence is reversed in the chromosome. The inversion sizes can range from tens to millions of basepairs[45].

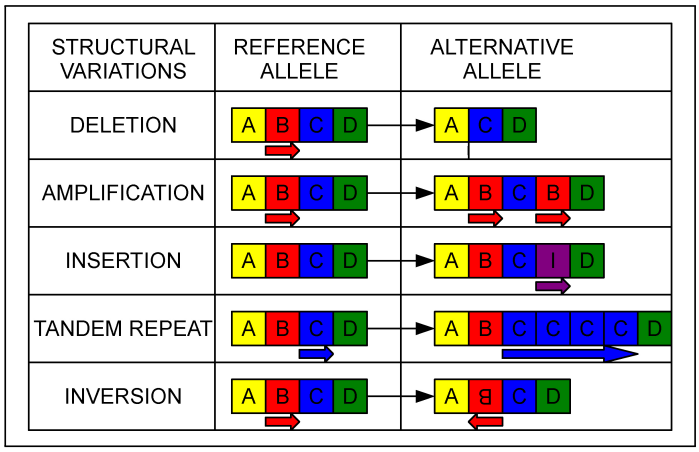


Figure 1.8: *Structural variations.* This figure shows the different structural variations that can be found in the genome. The arrows show the differences between the reference and alternative alleles.

1.2.3 Linkage disequilibrium and genotyping microarrays

The alleles that lie close in a chromosome sequence tend to be inherited together more often than distant alleles or alleles in different chromosomes[46]. The origin of this allele correlation, or linkage disequilibrium (LD), is the DNA recombination that takes place during cell meiosis. During gametogenesis, the chromosome pairs of each parent are mixed (i.e. chromosomal recombination), leading to new chromosome sequences. This stage of the reproductive process is crucial in increasing the genetic diversity of the next generation of individuals. Recombination occurs at random in the chromosome, although in certain regions, called hotspots, the probability is much higher[47]. Consequently, the closer two alleles are in a chromosome, the lower the probability that a recombination event will occur, and therefore, the higher the probability they will be inherited together. These groups of highly correlated alleles are also known as haplotypes[48]. Over multiple generations, the original haplotype block will be shortened by successive recombination events (Figure 1.9).

Consequently, only alleles that are very close together will have strong LD. LD is a powerful statistical measure that can be used to determine the time at which a certain mutation appeared and to study the genetic history of populations, including processes of negative and positive selection.

Although several different LD measures have been described (D, D', LOD)[49], r^2 is the most commonly used measure to define LD in genomic studies. r^2, also known as correlation coefficient, ranges from 0 to 1. When $r^2 = 1$ the two SNPs are in perfect LD (the genotype of one of the SNPs can be perfectly predicted based on the genotype at the other SNP).

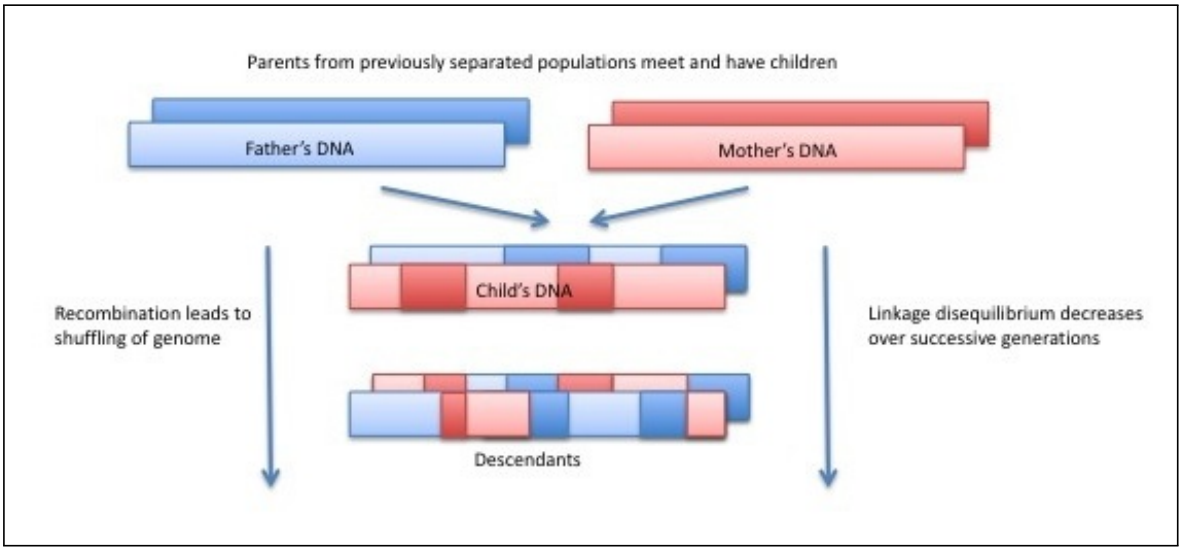


Figure 1.9: *DNA recombination.* This figure shows DNA recombination over successive generations leading to shorter haplotype blocks. Source: www.emory.edu.

Importantly, LD patterns in the human genome can be exploited to enable genetic association studies at the genome-wide scale. Although there are more than 40 millions of SNPs, the correlation patterns existing between these markers (i.e. haplotype blocks) allow the use of indirect analysis approaches. That is, for a certain LD block, it is not necessary to genotype all SNPs in this region to detect an association with a trait, since all variants are highly correlated, and the information can be captured by one or few SNPs (Figure 1.10)[50]. These SNPs, known as tagSNPs, are the basis of most published genetic studies in complex diseases and of all Genome-Wide Assciation Studies (GWAS). The International HapMap project (*www.hapmap.org*[51]) is an international collaboration project, that was created to characterize the main LD patterns present in the human genome, in order to enable large-scale genetic studies of human traits[52]. The data released by the HapMap project was crucial for the design genotyping microarrays. With only 500,000 selected markers, microarrays are able to capture most of the common genetic variation in the human genome.

More recently, the 1000 Genomes Project (1KGP, *www.1000genomes.org*)[53] has incorporated the new generation of sequencing technologies (i.e. Next-Gen), to produce genome-wide sequencing data for more than 2,000 individuals. This international collaboration

project, has therefore provided therefore an even more exhaustive mapping of the LD patterns in multiple populations, as well as including less frequent genetic variants (1% < MAF < 5%).

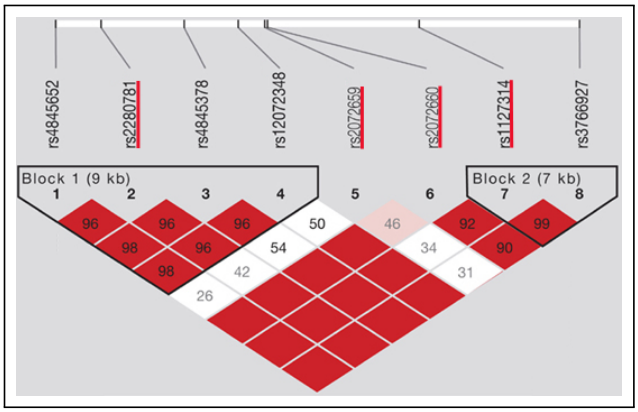


Figure 1.10: *TagSNPs and linkage disequilibrium.* This figure shows the correlation patterns (i.e. LD) between adjacent SNPs. The identified haplotypes based on these correlations are enclosed in black triangles. By selecting a reduced number of SNPs (tagSNPs -underlined in red-) of each haplotype, most of the genetic variation information in a genomic region can be captured.

The optimization criterion of tagSNP selection is the maximization of genomic coverage, defined as the number of SNPs that are in high LD (i.e. $r^2 > 0.8$) with any of the selected tagSNPs (Figure 1.11). Since LD patterns highly depend on the studied population, the selection of tagSNPs and the resulting genomic coverage might vary according to the analyzed population.

To date, microarray genotyping platforms can genotype millions of SNPs for each individual. The increasing drop in the costs of these genotyping technologies, has contributed to the explosion of GWAS of human traits.

High-density genotyping microarrays have been commercialized by several companies like Perlegen[54], Affymetrix[55] or Illumina[56]. Here we will focus on the Infinium-based beadchips developed by Illumina (Palo Alto, California, USA). Due to its superior quality and coverage[57], Illumina microarrays have been the platform of choice in most published GWAS in human traits (>70%). Also, Illumina was the genotyping platform of choice used in the *IMID-Kit* project (Figure 1.12).

The Illumina Infinium assays allow the multiplexed genotyping from 100,000 to 5,000,000 SNPs. Although the first microarray platforms were only based on common SNPs (MAF>5%), the most recent microarrays have incorporated the knowledge generated by the 1KGP to extend their coverage to more uncommon SNPs (2.5%<MAF<5%). As an example, the Omni1-Quad microarray provides genotyping on 1,138,747 markers and covers 94%, 93% and 78% of the common variation (MAF>5%) within the CEU (Utah residents of northern

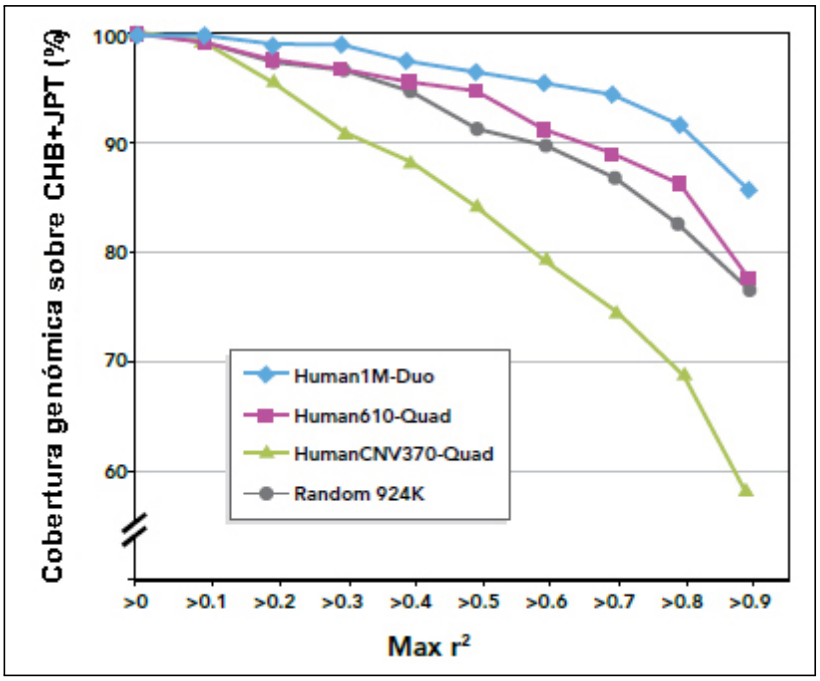


Figure 1.11: *Genomic coverage.* Comparison of the genomic coverage of multiple genotyping microarrays. An accurate selection of the tagSNPs allows to greatly increase the genomic coverage compared to a randomly selected set of SNPs. *Max* r^2 refers to the minimum LD with a tagSNP required to count a SNP as covered. Source: *www.illumina.com*.

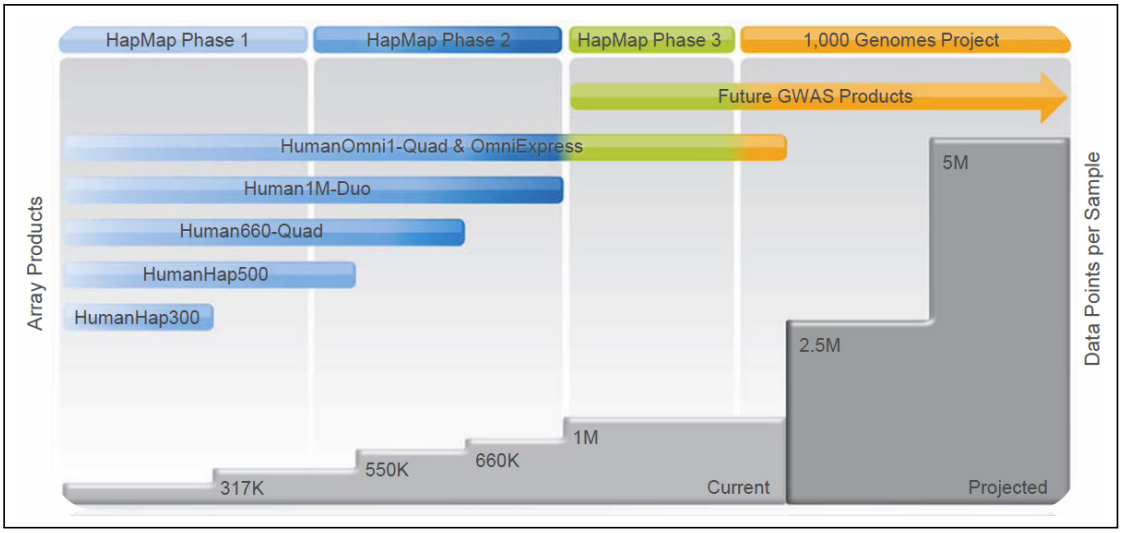


Figure 1.12: *Illumina BeadChip microarrays.* This figure shows the design and evolution of Illumina genotyping platforms: from the HumanHap300 that included 317,000 probes to the HumanOmni1-Quad that included probes targeting >1,000,000 different polymorphisms. Source: *www.illumina.com*.

and western European ancestry), CHB+JPT (combined panel of Han Chinese from Beijing and Japanese from Tokyo) and YRI (Yoruba from Ibadan -Nigeria-) populations, respectively.

The genotyping process of the Infinium assays is summarized as follows (Figure 1.13):

- Isolation and amplification: The genomic DNA from each individual sample is isolated (250-700 nanograms), amplified (x 1,000-3,000) and fragmented[58].

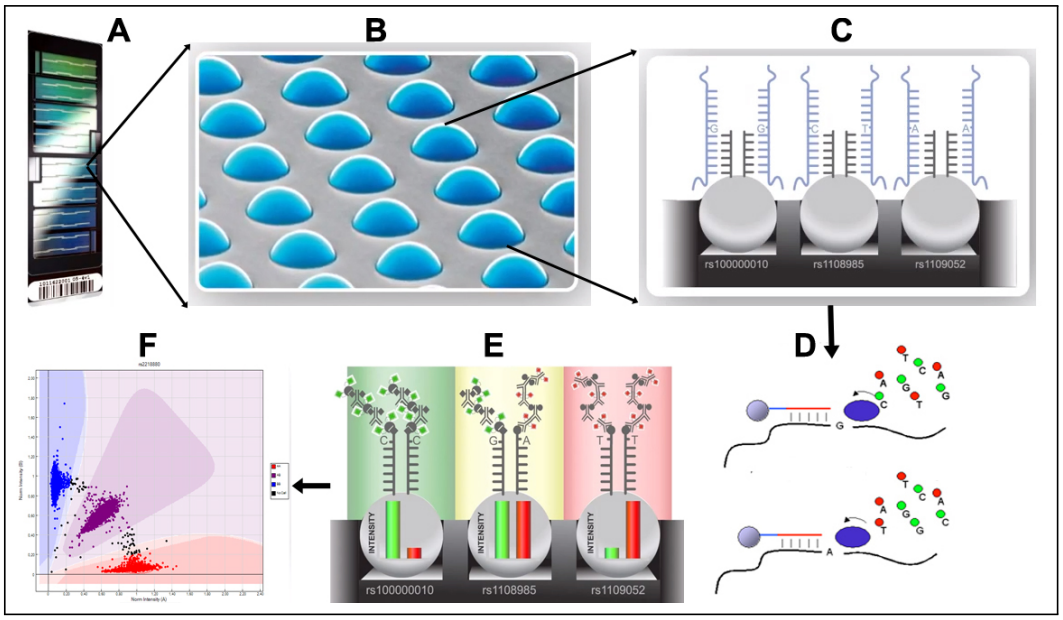


Figure 1.13: *Infinium assay.* (A) Genotyping microarray; (B) Silica beads where the DNA probes are fixed. Each SNP will have a probe pair, which will bind to each of the two possible alleles; (C) Hybridization of probes to their corresponding SNP alleles; (D) Nucleotide expansion of the bound probe to generate a specific fluorescent signal; (E) Probes bound to their corresponding alleles are scanned to detect the presence of each allele; (F) Finally, a genotyping algorithm used the observed fluorescent patterns from the probe pairs of multiple individuals to assign the corresponding genotypes. Source:*www.illumina.com*.

- Hybridization: The Illumina Infinium microarrays are designed with silica beads where the assay DNA probes are fixed (Figure 1.13B). The probes are 50 bp long and are designed to be complementary to the sequences of each SNP allele (i.e. two probes per SNP) (Figure 1.13C).

- Enzymatic simple base extension: After the target DNA segment is hybridized with the probe, the sequence is extended by a DNA polymerase enzyme with the use of labeled dideoxynucleotides (Figure 1.13C). Once linked to the primer, these dideoxynucleotides act as chain finishers and DNA polymerase stops adding nucleotides. The dideoxynucleotides corresponding to each one of the SNP alleles are labeled with different fluorochromes, thereby allowing the identification of each different allele with the use of high resolution scanners (Figure 1.13E).

- Genotype calling: The fluorescence at each probe pair generates different intensity clusters, according to the SNP genotype of each individual. Genotyping algorithms apply clustering methods to analyze at the same time the intensity data generated from multiple samples. The fluorescence intensities of each channel (i.e. defined as X and Y channels that correspond to each one of the SNP alleles) are processed to model the corresponding clusters of each one of the three possible SNP genotypes (Figure 1.13F).

SNP clusters are characterized by the ratio between the intensities of both channels: $I_X > I_Y$ homozygote samples for the first SNP allele, $I_X < I_Y$ homozygote samples for the second SNP allele and $I_X \simeq I_Y$ heterozygote samples. Additionally, the intensity data can also be used for CNV genotyping. If a probe targets a CNV, the variation in the number of copies of the DNA segment leads to an increase or decrease of the measured fluorescent intensities. Consequently, samples will also cluster at the total intensity level depending on the total number of copies they carry.

1.2.4 Genome-wide association studies

GWASs are based on the simultaneous analysis of thousands of SNPs to identify genetic loci associated to a trait of interest (Figure 1.14). During the last decade, GWASs have been highly successful in the identification of new biological pathways associated to the risk to develop diseases[19]. Despite its enormous potential, one crucial aspect for the success of GWAS in common complex diseases, is the accurate phenotype definition and selective collection of the genotyped individuals. Only with well-defined patient (and control) cohorts, it is possible to characterize the genetic architecture of human diseases. The identification of disease risk loci by GWAS has had two major impacts in medicine: first, the use of genetic testing to predict disease outcomes or to guide disease therapy and, second, the identification of new biological mechanisms associated with disease physiopathology that can lead to the development of new, more efficient, drugs[19;57].

There are two main types of GWAS designs: family-based and population-based studies. The former is based in the analysis of related individuals such as parent-offspring trios[60]. Most of the methods for analyzing family-based data use the transmission of alleles to the offspring to assess the evidence of genetic association[60]. Conversely, in population-based studies a large cohort of unrelated individuals is genotyped, and the distribution of allelic frequencies according to the phenotypic variable is used to assess genetic association (Figure 1.15). Although family-based studies are robust against spurious associations due to ethnicity, admixture or population stratification[61], population-based studies can be more powerful to identify risk variants since, generally, much larger samples of individuals can be collected[62;63]. To date, most published GWAS are population-based[60].

GWASs can be used to analyze two primary classes of phenotypes, either quantitative or categorical. Quantitative traits are usually preferred since they can have more statistical power to detect genetic effects[65], and respond to the polygenic nature of common and complex diseases. In many occasions, however, quantitative traits are not available. In this case, individuals are usually classified as affected or unaffected, and analyzed using a case-control design[57]. The association of each SNP to the studied trait can be analyzed

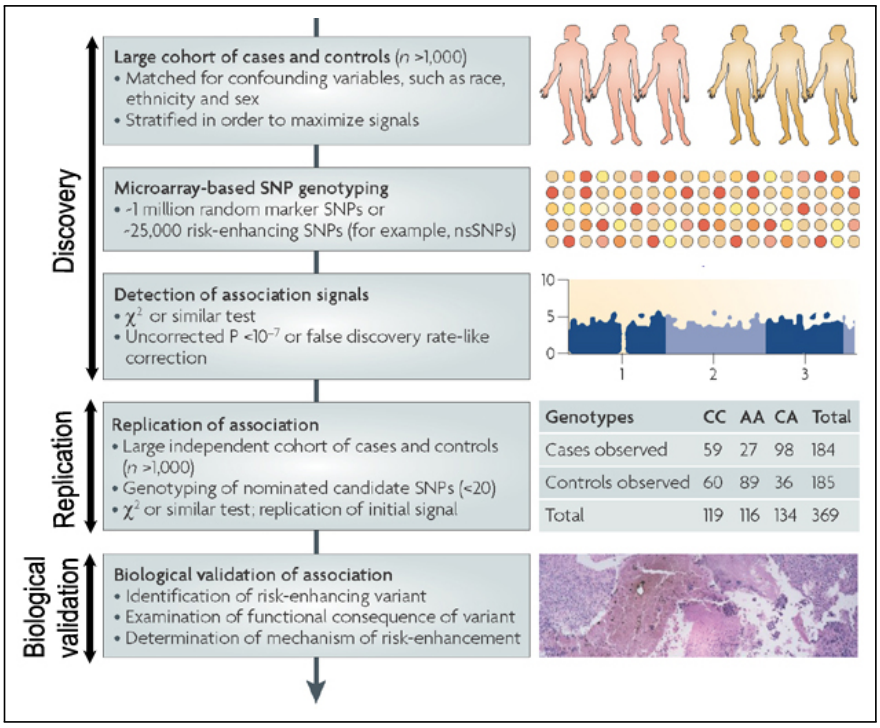


Figure 1.14: *Genome-wide association studies.* GWAS studies are commonly performed in two stages. During the study design stage, the phenotypes of interest are defined, as well as the potential confounding variables (i.e. ethnicity, sex and smoking behaviour). The individuals of the selected cohort are then genotyped using genotyping microarrays. The resulting genotyping data is subsequently used to perform the statistical tests for each genomic variant. Once the regions in the genome that are more significantly associated with the trait of interest are identified, they are tested in an independent (validation) cohort. The replication step ensures that the observed association in the discovery stage is not a false positive. Finally, the functional impact of the associated genomic variants may be further studied to explain how these variants are linked to the studied phenotype. Source: Adapted from *Kingsmore S et al*[59].

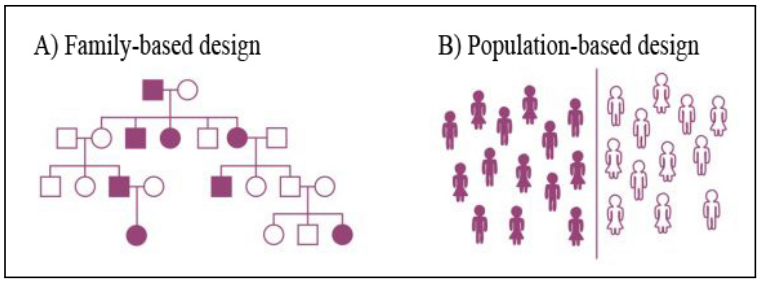


Figure 1.15: *Population-based and family-based GWAS design.* Family-based designs are based on the co-segregation of alleles and a trait within a family, while population-based designs are based on the differences in allele frequencies according to the studied phenotype (i.e. case-control). Source: Adapted from *Dick D et al*[64].

under different genetic association models. Each model implies different assumptions on how the different variant alleles contribute to the risk (i.e. recessive, dominant, genotypic, and allelic; Figure 1.16)[66;67]. The statistical tests used to analyze the association between genetic variants and different traits will depend on the phenotype class: quantitative traits are usually analyzed using linear regression or Analysis of Variance, while categorical or

binary traits are usually analyzed using contingency table analysis methods (e.g. chi-square test) or logistic regression[57;67].

Table 2. Genotypes, alleles, and statistical data of rs7903146 SNP at the TCF7L2 in diabetic and control subjects

Genotype distribution				
	CC	CT	TT	Total
T2D	49 (43.4%)	47 (41.6%)	17 (15%)	113
Controls	70 (50.4%)	63 (45.6%)	6 (4.3%)	139
Total	119	110	23	252
Allelic distribution				
	C		T	Total
T2D	145 (64.2%)		81 (35.8%)	226
Controls	203 (73%)		75 (27%)	278
Total	348		156	504
Genetic models and statistical data				
Genetic Model	Odds ratio (95% CI)	Chi-square	Degrees of freedom	P
Association (CC *vs.* CT *vs.* TT)		8.70	2	0.012
Dominant (CT + TT *vs.* CC)	1.32 (0.80-2.18)	1.22	1	0.260
Recessive (TT *vs.* CT + CC)	3.92 (1.49-10.3)	8.64	1	0.003
Co-dominant (CT *vs.* CC + TT)	1.16 (0.70-1.92)	0.35	1	0.55
Homozigote (CC *vs.* TT)	4.04 (1.48-11.0)	8.30	1	0.004
Heterozigote (CC *vs.* CT)	1.06 (0.63-1.80)	0.06	1	0.81
Allele (C *vs.* T)	1.51 (1.03-2.21)	4.58	1	0.032

Figure 1.16: *Genetic association models.* This figure shows an example of the distribution of genotypic and allelic data from a SNP (rs7903146) in the control and case groups of a Type 2 Diabetes GWAS[68]. Between the genetic models tested, the recessive model obtains the largest significance. Source: *Barcelos G et al*[68].

In addition to the selection of the most appropriate genetic model of association and the statistical analysis test, another important consideration in GWAS analysis is the multiple testing problem (Figure 1.17). GWAS involve the analysis of thousands to millions of SNP tests and, therefore, the probability of finding a statistically significant result by chance (i.e., false positive) is very high.

Several methods have been proposed to adjust for multiple testing in GWAS: Bonferroni and Sidak corrections are simple methods that adjust for multiple testing, permutation testing has been established as a gold standard for correcting association values[67]. Nevertheless, the scientific community has increasingly acccepted $\alpha = 5 \cdot 10^{-8}$ as a genome-wide threshold for statistical significance in GWAS[67;69]. This means that all SNPs in a GWAS that show a *P*-Value lower than $\alpha = 5 \cdot 10{-}8$ can be considered robust genetic associations to the trait of interest.

In the typical GWAS approach, the set of SNPs that are selected for replication depend on the level of statistical significance in the genome-wide analysis stage. However, recent approaches also consider SNPs with significance association values under the genome-wide level of significance. These GWAS prioritization approaches[70;71] use scoring methods that also integrate the information of biological pathways and the functional role and other biological features of the variants.

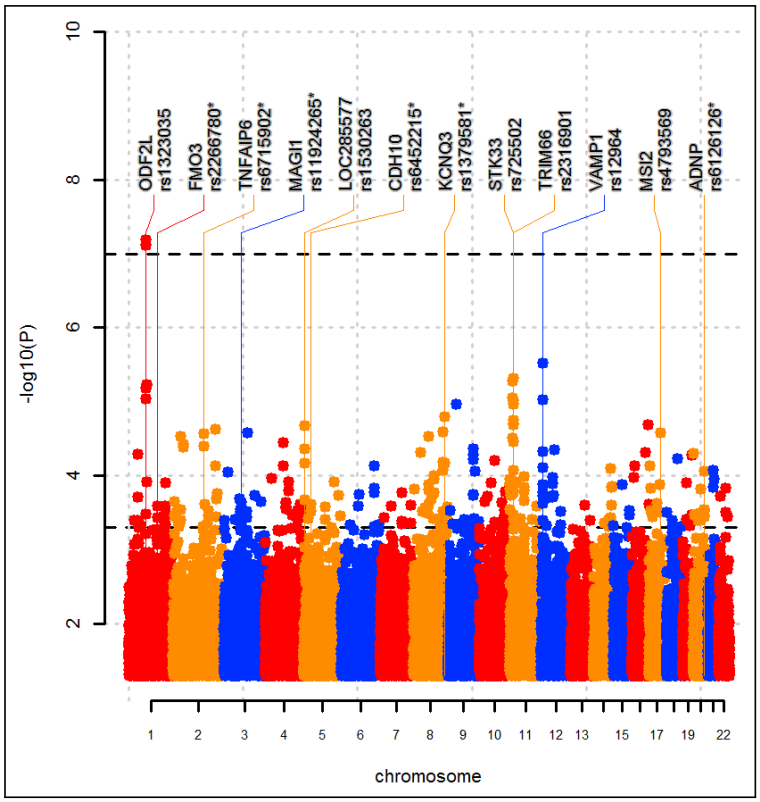


Figure 1.17: *Manhattan plot.* This type of graphic representation is usually used to show the association results in GWASs. In this figure, each point represents a genetic variant. The variants are plotted according to their genomic position (bp in the chromosome, each color represents a different chromosome) and the y-axis shows their level of statistical association to the studied trait ($-log_{10}(P-Value)$).

1.2.5 Genotyping algorithms

SNP genotyping algorithms

There are a large number of methods available to process microarray intensities and infer the genotypes of each SNP in a given set of samples (Figure 1.18). Three of the most commonly used algorithms in Illumina microarrays are GenCall[72], GenoSNP[73] and M3[74]. These algorithms use the intensities of X and Y channels generated by the microarray for each SNP assay and each sample. The main differences in the genotyping algorithms are based on the intensity normalization procedure and on the type of models used for genotype clustering (Figure1.18).

- **GenCall**[72]: GenCall is the genotyping method provided by the microarray manufacturer (Illumina). In this method, intensities are normalized using an affine transformation that rotates and scales channel X and Y intensities reducing channel cross-correlation[75]. This normalization procedure is performed separately for each bead pool. Bead pools are assay groups that have been built together and are also located

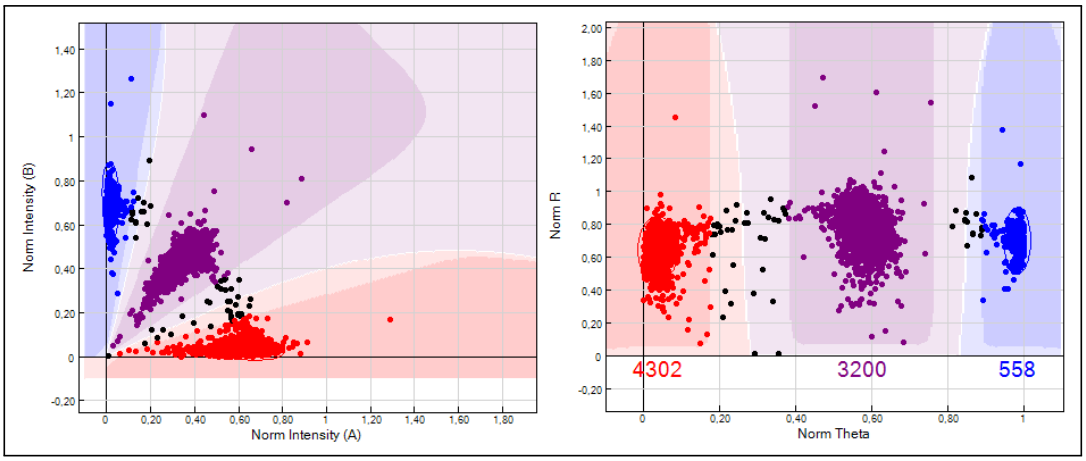


Figure 1.18: *SNP genotyping.* This figure shows the intensities generated by one SNP assay using Cartesian (left) and polar (right) coordinates. Each point corresponds to one sample and represents the measured intensities for each allele. The objective of genotyping algorithms is to identify the clusters corresponding to each genotype (AA, AB and BB). These clusters are mainly characterized by the intensity ratio between channels A and B (i.e. polar coordinate θ).

close together in the BeadChip and therefore, they require specific normalizations. Normalized intensities are then transformed to polar coordinates (R and θ) to perform the clustering. Using default centroid positions predefined for each assay and a neural networks algorithm, this method re-estimates the centroids of each cluster based on the data available for each probe and assigns a genotype to each sample based on the new cluster estimates.

- **GenoSNP**[73]: This genotyping method (University of Oxford, USA) is characterized by using a sample-based clustering, where each sample is independently genotyped using the information of all the available assays. For each sample, GenoSNP first normalizes X and Y intensities by each bead pool. Once normalized, it fits a four-component mixture model to the log-transformed intensities. This model jointly fits all the intensities coming from assays of the same bead pool.

- **M3**[74]: This method (Yale University, USA) is a two-stage genotyping algorithm. First, it uses an assay-based approach that fits the genotype clusters of each assay. Second, it identifies subsets of SNPs with low MAFs (i.e. <0.05) and low genotyping quality scores to improve their genotyping. To improve the cluster definition of these SNPs, M3 usesthe reference clusters of high-quality SNPs with similar intensity distributions as a prior cluster model for the low-quality SNPs.

CNV genotyping algorithms

While SNP genotyping is based on clustering at the intensity ratio level ($I_A > I_B \rightarrow AA$, $I_B > I_A \rightarrow BB$ and $I_A \simeq I_B \rightarrow AB$), CNV genotyping is based on differences at the global

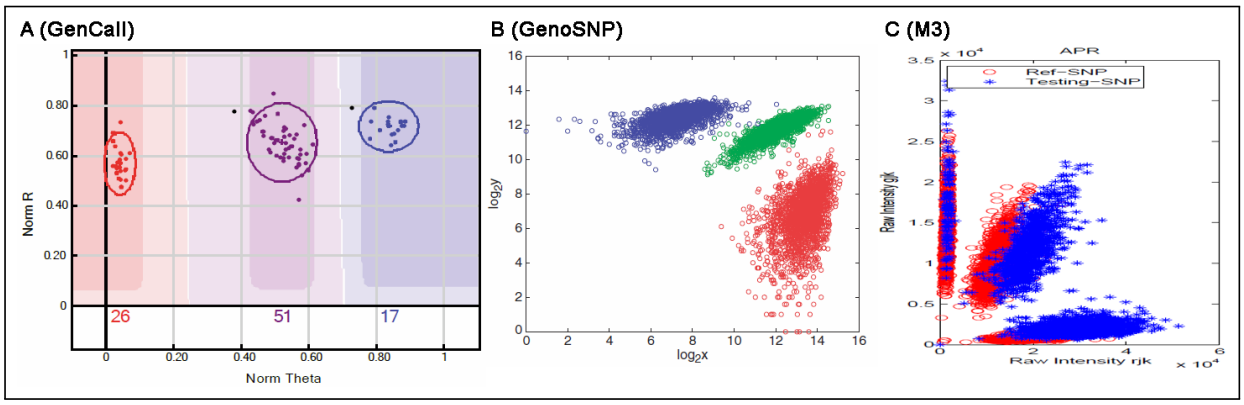


Figure 1.19: *SNP genotyping methods.* (A) Polar intensities of one assay where each point corresponds to one sample. GenCall assigns a genotype (i.e. color coded) using neural networks and prior knowledge. (B) Log-normalized intensities used by GenoSNP. The points correspond to different probes located in the same bead pool and refer to only one sample. (C) Re-genotyped SNP (blue) using information from a reference SNP (red) in M3 method.

intensity level. The global intensity level measured by the microarray is proportional to the number of DNA copies.

Two of the most commonly used algorithms for CNV genotyping using Illumina microarrays are PennCNV[76] and QuantiSNP[77]. Both methods use sample-based approaches that analyze the changes in total intensity measurements across adjacent genomic assays. These methods use two features that summarize sample intensities in relation to a reference status. These features are derived from the polar coordinates R and θ computed from the corresponding Cartesian coordinates I_A and I_B (Figure 1.20):

- Log R Ratio: To compute this feature, a regression line is fitted through the centroids of the three SNP genotyping clusters. This line provides the expected value of R ($\widetilde{R}_n$) given a θ_n value. The Log R Ratio is defined as the logarithm of the ratio between the measured and the expected R values: $LogRR = \log \frac{R_n}{\widetilde{R}_n}$. Since the most common state (i.e. reference) is 2 copies (i.e. diploid), hemizygous deletions (i.e. 1 copy) are expected to have $LogRR = \log \frac{0.5\widetilde{R}_n}{\widetilde{R}_n} = -1$ and amplifications (i.e. 3 copies) $LogRR = \log \frac{1.5\widetilde{R}_n}{\widetilde{R}_n} = 0.60$.
- Allelic intensity ratio: The θ values corresponding to the centroid of each SNP genotyping cluster are assigned to 0, 0.5 and 1. From these values, the θ_n values of each sample are interpolated within the range [0,1].

Using Log R Ratio and allelic intensity ratio, PennCNV and QuantiSNP apply hidden Markov models (HMMs) to stratify genomic segments according to their estimated number of copies. The HMMs model assumes that the observed intensities are related with an unknown copy

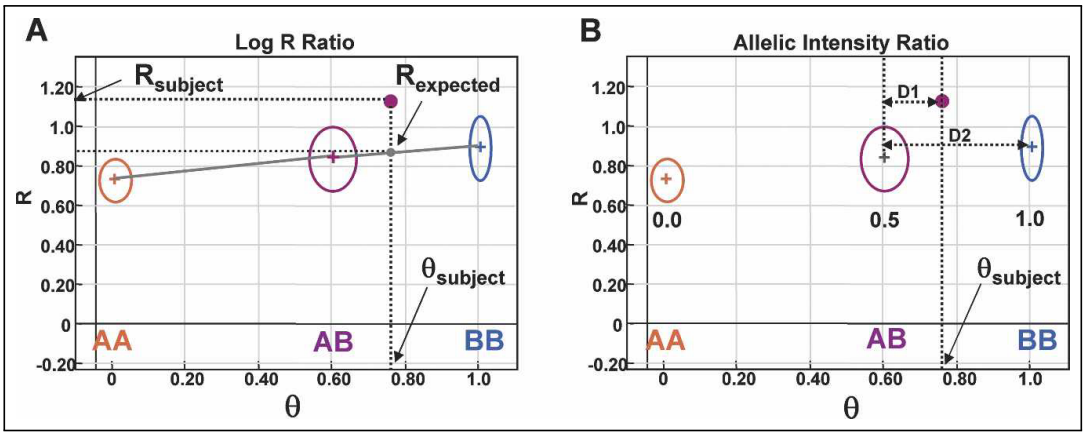


Figure 1.20: *Allelic frequency and LogRratio.* (A) The logRratio compares the observed intensity (Rsubject) of the sample to the expected intensity (Rexpected; gray dot) based on the observed allelic ratio. (B) The genotype clusters (circles) are used to convert θ values to B allele frequency. Source: *Peiffer D et al*[75].

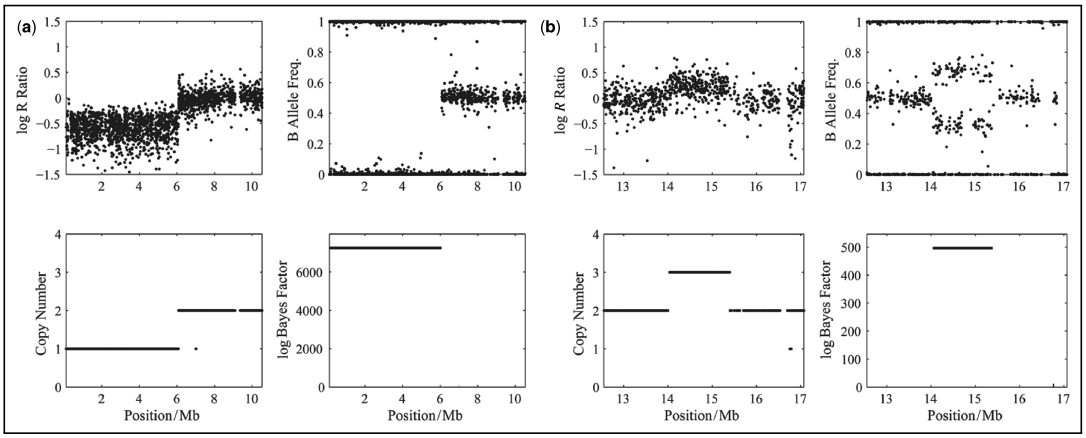


Figure 1.21: *CNV detection.* This figure shows the expected LogRratio and allelic frequency values expected in the case of deletion (A) and amplification (B). Using this data (upper panels) the algorithms compute the number of copies per sample (lower panels). Source: *Colella S et al*[77].

number state through emission distributions. HMMs also assume that the number of copies in a single assay is highly correlated with adjacent assays (Figure 1.21).

Both PennCNV and QuantiSNP are based on a per-sample approach and use summarized measurements relative to a reference set of samples. This type of approach has proven to work well to genotype large CNVs. However, this approach is highly inefficient at identifying short size CNVs. Also, these two approaches do not use the information generated by the simultaneous analysis of multiple individuals, leading to high false negative rates, particularly with small CNVs[78]. Therefore, there is a need for new methods that can use the information generated by considering all samples together in order to improve the detection of small CNVs in GWAS.

1.3 Metabolomics

Note: Some sections of this chapter have been extracted from the article (Appendix D.1):
A. Alonso, S. Marsal and A. Julià. **Analytical methods in untargeted metabolomics: state of the art in 2015**. Front. Bioeng. Biotechnol, 2015, 3:23.
DOI: 10.3389/fbioe.2015.00023.

Metabolites are the intermediates and end products of multiple enzymatic reactions and, therefore, they are the most informative proxies of the biochemical activity in an organism. Metabolomics is the study of the metabolite composition of a cell type, tissue, or biological fluid. To date, metabolomics is one of the "omics" approaches that is most contributing to the progress of challenging research areas like the personalization of treatments in medical practice.

The current metabolomic analysis technologies have enabled the characterization of the complete set of metabolites - the metabolome- in multiple types of biological samples[79]. These applications cover diverse research areas like plant biology[80], nutrition[81;82], animal breeding[83], drug discovery[84;85], and the study of human diseases[86;87]. The biomedical field is actually one of the most active areas of development in metabolomics, and includes the search for diagnostic and prognostic biomarkers as well as predictors of treatment response[88–90].

To date, the two main technical approaches for the generation of metabolomic data are nuclear magnetic resonance (NMR) and mass spectrometry (MS;[91]). NMR is a fast and highly reproducible spectroscopic technique that is based on the energy absorption and re-emission of the atom nuclei of the sample molecules due to variations in an external magnetic field (Figure 1.22)[92]. Depending on the atom nuclei being targeted by the applied magnetic field, different types of metabolomic data are generated. In biomedical studies, hydrogen is the most commonly targeted nucleus (^{1}H-NMR), due to its natural abundance in biological samples. Although less frequent, other atoms like carbon (^{13}C-NMR) and phosphorus (^{31}P-NMR) are also targeted by NMR, providing additional information on specific metabolite types[93].

The resulting spectral data in NMR not only allows the quantification of the concentration of metabolites, but it also provides information about its chemical structure. The spectral peak areas generated by each molecule are used as an indirect measure of the quantity of the metabolite in the sample whereas the pattern of spectral peaks generated by the molecule is used to identify the type of metabolite. The spectral data obtained with NMR techniques can be one dimensional or two dimensional. One dimensional NMR (1D-NMR) spectra are based on a single frequency axis, where the peaks of each molecule are placed within

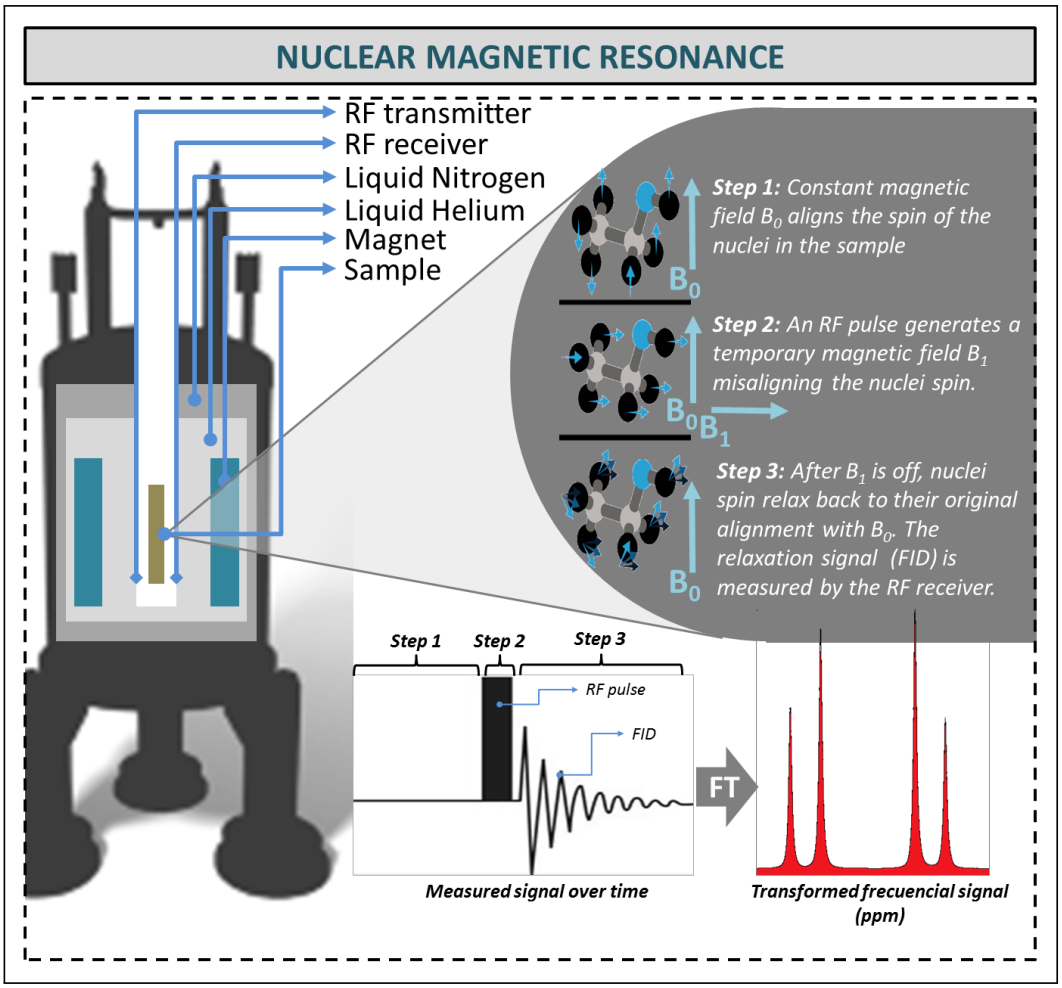


Figure 1.22: *Nuclear magnetic resonance.* The NMR spectral acquisition is based on the behavior of the molecule spin under magnetic field variations. First, a constant magnetic field is applied to the sample, aligning the spins of all their molecules (i.e. step 1). The next step consists of applying a RF pulse to generate an interfering magnetic field which temporary misalign the molecules' spins (i.e. step 2). Once the interfering magnetic field disappears (i.e. step 3), the molecules' spins relax back to their original alignment. This spin relaxation results in a signal, FID, which can be measured and, after applying on the FT, is transformed on a peak spectrum where each peak is characterized by its amplitude (vertical axis) and its chemical shift (horizontal axis). The latter is usually measured in ppm and refers to the difference between the resonance frequency and that of a reference substance divided by the frequency of the spectrometer.
FID: Free induction decay; NMR: Nuclear magnetic resonance; ppm: Parts per million; RF: Radio frequency; FT: Fourier transform.

its resonant frequencies (Figure 1.23)[94]. 1D-NMR is the most commonly used method in high-throughput metabolomics studies. Two dimensional NMR (2D-NMR) spectra are based on two frequency axis, and its use is generally restricted to the characterization of those compounds that cannot be directly identified with 1D-NMR spectra. The second dimension in 2D-NMR improves peak separation and gives additional and important information on the chemical properties of the metabolite[95]. 1D- and 2D-NMR frequency axes are usually referenced by the chemical shift expressed in parts per million (ppm). The chemical shift

is calculated as the difference between the resonance frequency and that of a reference substance, subsequently divided by the operating frequency of the spectrometer[96].

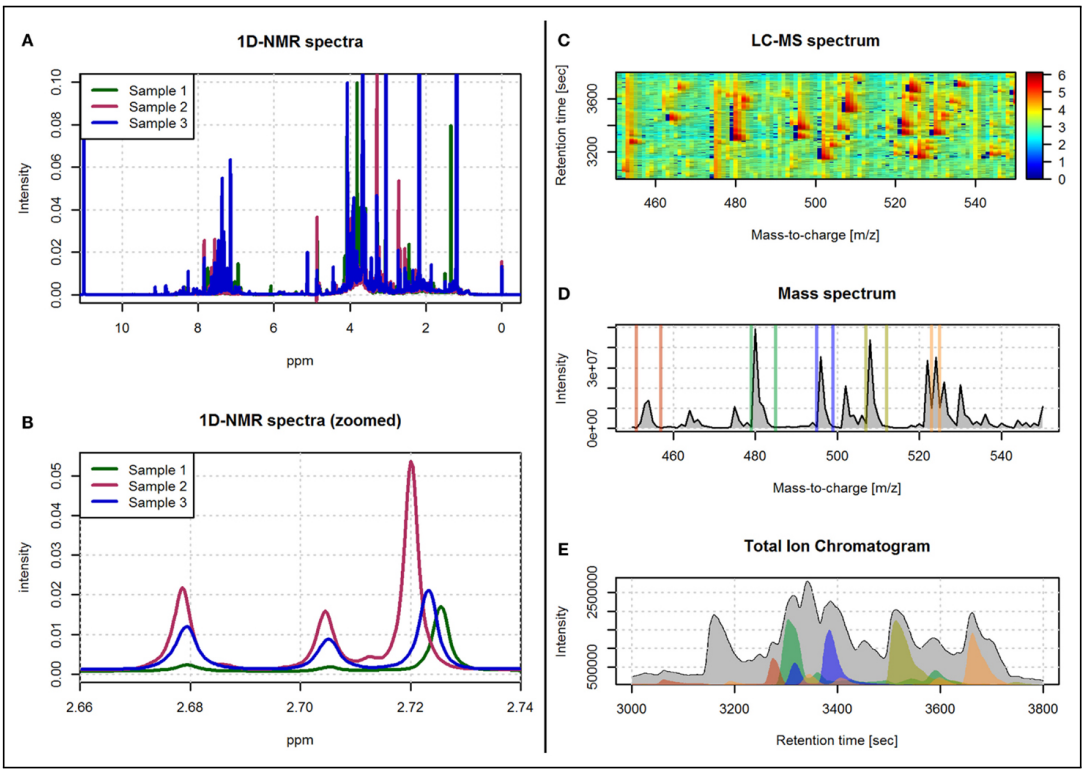


Figure 1.23: *Examples of spectra obtained with ^{1}H-NMR and LC-MS technologies.* (A) An example of three spectra obtained with 1D ^{1}H-NMR. (B) A zoomed view of the spectra in (A) in the 2.66-2.74 ppm range. (C) An example of a LC-MS spectrum with color-coded intensity and referred by the m/z and retention time axes. (D) The sum of the LC-MS spectrum across the m/z axis. (E) The total ion chromatogram (i.e. sum of the LC-MS spectrum across the retention time axis). The colored regions in (E) correspond to the sum of the LC-MS spectrum limited to the m/z ranges depicted with the same color in (D).

Untargeted metabolomic studies are characterized by the simultaneous measurement of a large number of metabolites from each sample. This strategy, known as a top-down approach, avoids the need to define a prior specific hypothesis on a particular set of metabolites and, instead, analyses the global metabolomic profile. Consequently, these studies are characterized by the generation of large amounts of data. This data is not only characterized by its volume but also by its complexity and, therefore, there is a need for high performance bioinformatic tools that can efficiently extract and analyze the relevant biological information.

Figure 1.24 shows the prototypic methodological pipeline of an untargeted metabolomic study. This methodological pipeline starts with the processing of the spectral data to generate the sample metabolic information (i.e. metabolic features). The different methods available to process spectral data are revised in subsection 1.3.1. Together with metabolite-identification methods, spectral processing methods are highly dependent on the analytical

technique used (e.g., NMR, LC-MS, or GC-MS). Once the complete set of metabolomic features has been generated, univariant and multivariant data analysis methods can be then applied to investigate: (a) the general structure of the metabolomics data in the dataset and (b) how the different metabolic features are related with the phenotypic data associated with the samples. In subsection 1.3.2 we address the important technical issue that is the identification of the metabolites underlying the observed metabolic features (i.e. peak areas and spectral bins).

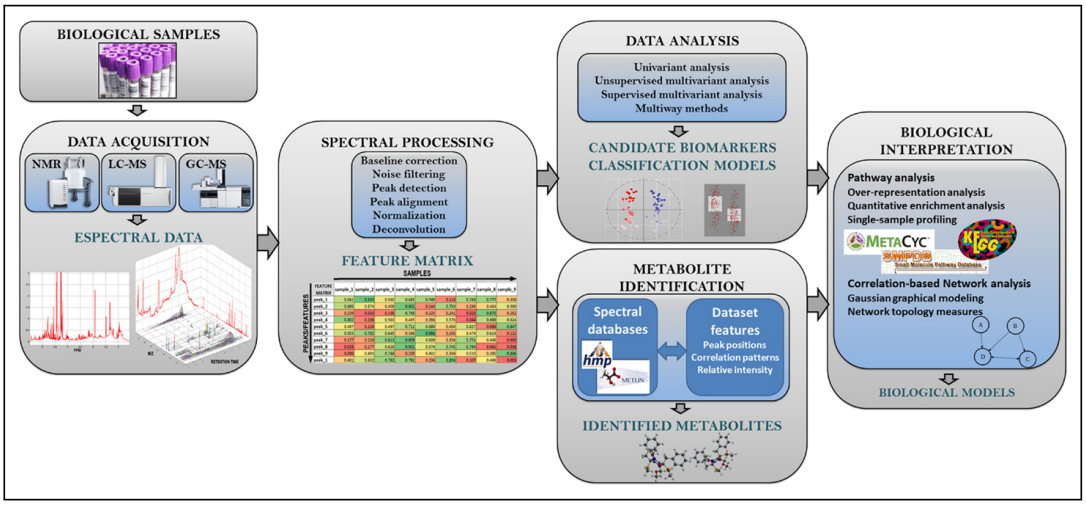


Figure 1.24: *Analysis workflow in untargeted metabolomic studies.* This figure shows the different steps of the metabolomic analysis pipeline. FOCUS, the bioinformatics tool developed during this Doctoral Thesis covers both *Spectral processing* and *Metabolite identification* steps.

1.3.1 Spectral Processing

Spectral processing is a methodological approach that intends to accurately identify and quantify the features in the sample spectra of a metabolomics study (Figure 1.25). In high-throughput metabolomics studies, spectra are sequentially or jointly processed to obtain a final set of feature quantifications. Spectral processing is also necessary to ensure that each final measurement will refer to the same metabolite in all samples. The data resulting from spectral processing is generally arranged in a feature quantification matrix (FQM) that contains the quantification of the metabolic features of all the analyzed samples, and that will be used as input for subsequent statistical analysis.

Spectral Pre-Processing

In order to improve the signal quality and reduce possible biases present in the raw data, several pre-processing steps are usually applied. In NMR-based spectra, baseline correction is used to remove low frequency artifacts and differences between samples that are generated by experimental and instrumental variation[97–100]. After this, the application of

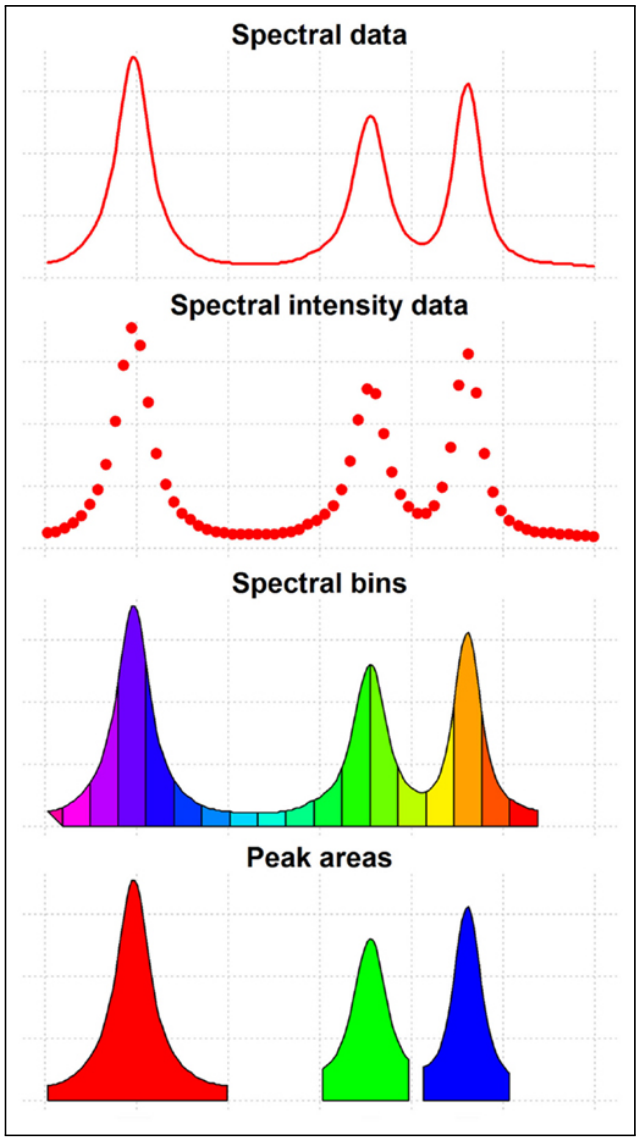


Figure 1.25: *Features of spectral data.* This figure shows the three different types of features that can be extracted from spectral data and used for data analysis. First, statistical analysis can be performed using all the data points of the spectrum (i.e. spectral intensity data); second, integrating evenly spaced regions of the spectrum (i.e. spectral bins) and, third, automatically detecting the spectral peaks and integrating their corresponding areas.

high-frequency filters may be necessary to remove the electronic noise present in the data that is generated by the measurement equipment[101].

Feature-Detection

The objective of the feature-detection step is to identify and quantify the features present in the spectra. Peak-based methods detect the peaks across the spectrum and integrate their areas to provide a quantification of the underlying metabolite[102–104]. In this approach, spectral alignment is also generally applied either before or after peak detection. In NMR studies, binning-based approaches have been commonly used to detect feature peaks in

complex biological samples[105]. However, these methods perform poorly compared to peak-based methods, particularly in those cases where there is significant spectral unalignment, or in those cases where multiple peaks from different metabolites are captured by the same spectral bin[106]. For these reasons, peak-based methods are increasingly being used in NMR-based studies[105]. Nonetheless, there have been recent developments in binning algorithms, particularly in the detection of the optimal binning boundaries, that have improved the performance of this feature-detection approach[107].

Peak overlap is also a common problem in NMR-based studies. Overlapping peaks are treated as one same feature both in binning and peak-based approaches[108]. Consequently, the results obtained from the analysis of these variables can be often hard to interpret since the contribution of each peak to the total area is unknown. In order to attempt to solve this problem, spectral deconvolution methods have been developed[109]. These methods, which are based on the fitting to metabolite spectral templates, are able to extract independent metabolite quantifications from a set of overlapping peaks. The main disadvantage of this type of algorithms, however, is that they depend on the existence of spectral libraries of each metabolite and, therefore, they are unable to quantify peaks arising from previously uncharacterized metabolites.

Two main procedures are performed for feature detection: peak detection and spectral alignment.

- **Peak detection:** The most commonly used peak detection algorithms analyze each sample spectrum independently[110–112]. These methods are based on two analytical steps[113]. In the first step, the spectra are smoothed. For this objective, multiple different filters are available (i.e. moving average, Gaussian, Savitzky-Golay...;[113]). From these, the Wavelet transform-based filters have demonstrated a superior performance, although at the expense of a higher computation time[110;114]. The performance improvement is mainly due to the ability of the Wavelet transform to work with unequal peak widths, which are characteristic of metabolomic spectra. In the second step, the different metabolite peaks are identified using one or multiple detection thresholds. These thresholds are applied to different parameters such as the signal-to-noise ratio, the intensity, or the area of each peak from the resulting filtered spectra[113]. In metabolomics studies involving large numbers of samples, a frequency filter (i.e. consensus peak signal), can be also applied so that only those peaks that are present in a minimum percentage of samples are selected for downstream analysis.

- **Spectral alignment:** Spectral alignment is one of the main processing steps in metabolomic studies involving multiple samples. When analyzing multiple spectra,

the position of the peaks corresponding to the same metabolic feature may be affected by non-linear shifts. In NMR-based studies, these shifts are observed in the ppm axis and are usually introduced by differences in the chemical environment of the sample like ionic strength, pH, or protein content[108;115]. In MS-based studies, peak shifts are mainly observed across the retention time axis, and are generally associated with changes in the stationary phase of the chromatographic column[116]. Spectral alignment methods must be therefore applied to correct this undesired variability in the samples that can profoundly affect the quality of the study. Spectral alignment algorithms can be divided in two main groups: (i) spectral alignment methods, where the spectral data is aligned before peak detection and (ii) peak-based alignment methods, where spectral peaks are aligned across samples once they have been detected using their coordinates (ppm in NMR, and m/z and retention time in LC/GC-MS).

Spectral alignment methods are classified into warping and segmenting methods. Warping methods are based on the application of a non-linear transformation to the ppm (in NMR spectra) or the retention time (in LC/GC-MS) axis in order to maximize the correlation between the spectra. The alignment is then performed by either stretching or shrinking spectral segments to reach this correlation maximization. Among these methods, correlation optimized warping (COW) and dynamic time warping (DTW) are the most commonly used. COW is a segmental alignment method that aligns one sample spectrum toward a reference spectrum[117]. This is done by splitting the original sample and reference spectra into small segments, and by separately aligning each pair of segments. Alignment is performed through dynamic programing in such a way that limited changes in segment lengths are allowed. This way, the overall correlation between both spectra is effectively maximized. In the particular case of crowded spectral regions with large peak shifts, COW has demonstrated to perform particularly well compared to other methods. An alternative to COW method, DTW is a spectral alignment method[117] that is also based on dynamic programing, and where a warping path is computed to which the connected data points of each spectrum are equivalent. During this last decade, other warping approaches have been developed[118–121].

Spectral segmenting methods differ from spectral warping methods in that alignment is performed by applying a constant shift to all the spectral points. These methods either align the overall spectra or split the spectra into smaller segments and independently align each resulting segment. The Icoshift algorithm[122] is one of the most commonly used segmentation methods, and is based on the convergence toward a reference signal. This convergence is performed by applying shifts that maximize the segment spectral correlation, which is normally computed using the fast Fourier trans-

form (FFT) to speed up the required calculations[123]. Icoshift and other correlation-based methods can also be combined with automatic segmentation methods[124], which are able to optimally split the spectra in order to improve the alignment of the resulting spectral segments. However, the use of a reference spectrum has several disadvantages such as the fact that the reference spectrum may not be representative of the spectral diversity present in the samples or the biases produced the presence of multiple peaks in the same alignment window. Under these conditions, the methods based in correlation maximization are prone to align the most relevant peak of each sample regardless of whether they correspond to the same metabolic feature or not.

Fast Fourier transform-based segmenting methods not only are able to process large metabolomics datasets in a reduced amount of time, but also have shown to perform better than spectral warping methods[4;122;125;126]. Within the different segmenting methods, reference-free methods avoid the biases introduced by using reference spectra, but at a cost of being more computationally intensive.

Of relevance, the results reported by several performance comparison studies using either NMR or MS have demonstrated that spectral alignment algorithms have a good performance irrespective of the analytical technique that has been used (MS or NMR;[125;127]. Consequently, methods that were initially developed to align NMR spectra are also applied to align MS spectra and vice versa.

Compared to the warping and segmentation alignment methods, peak-based methods are applied after peak detection. In these methods, peak coordinates are used to perform the alignment. This type of method is implemented in the XCMS software[128], one of the most commonly used methods to process data from LC-MS studies. Given that the shifts along the m/z axis are minimal and the m/z axis has a high resolution, the data can be safely binned in m/z intervals, and peak alignment performed on each bin along the chromatographic time. The XCMS algorithm computes the retention time boundaries within which the observed peaks are expected to represent the same metabolomic feature across the different samples. The computation of these retention time boundaries is performed by using a kernel density estimator. Another common alignment method used in MS is the RANSAC algorithm[112]. In this approach, the corresponding peaks across samples are identified by using a LOESS regression on different retention times and m/z windows.

Feature Normalization

In order to perform an accurate quantification of the features in a metabolomic analysis, a data normalization step is generally required. The objective of normalization is to remove

undesired systematic biases, so that only biologically relevant differences are present in the data. This procedure is crucial when analyzing complex biofluids like blood, where the differences in metabolite concentration between samples can be high, and the introduction of internal standards is complicated. Although multiple statistical models have been developed for this objective[129;130], the two perhaps most commonly used methods are the use of endogenous stable metabolites (like creatinine in urine) and the use of the total spectral area (i.e. area under the curve (AUC);[108;131]).

1.3.2 Metabolite Identification and Spectral Databases

Metabolite identification is one of the major challenges of high-throughput metabolomic analysis. This step is indispensable to providea biological significanceto the associated features in a metabolomic study. In MS-based studies, the common metabolite-identification approach is based on querying metabolomic databases for the neutral molecular mass values of the identified peaks using a tolerance window. The neutral molecular mass is inferred from the peak m/z value, and depends on the chemical nature of the identified peak (i.e. ionization mode and ionization adduct). Assuming no prior knowledge, each peak m/z value can lead to multiple plausible neutral molecular masses that can represent different ionization adducts (i.e. H^+, Na^+, K^+). This multiplicity often results in a high number of false positive identifications. In order to reduce false positives, several methods have been developed. AStream and Camera are methods designed to identify isotopic and adduct patterns in order to reduce data complexity in MS experiments[132;133]. The PhD student was involved in the development of the former method before starting this thesis. Using these approaches, the chemical nature of each selected ion peak is estimated, and only one neutral mass is inferred from each identified pattern (Figure 1.26). Using these methods has the added advantage of improving the ascertainment of true biological compounds.

In NMR-based studies, automatic metabolite identification is commonly performed by matching the measured NMR peaks against a set of reference metabolite patterns. Each metabolite reference spectrum is defined by one or multiple peaks, which are characterized by their ppm positions and their relative intensities. MetaboHunter is an online tool for identifying compounds by matching the reference peak positions against the list of detected peak positions[134]. However, this approach can lead to high false positive rates, since it only uses one peak parameter to match reference peaks. The MetaboHunter approach has been superseded by more recent methods based on the valid cluster concept[135;136]. In addition to using the ppm position, these methods include peak intensities and inter-sample intensity correlation as parameters for matching data peaks to reference peaks. Nevertheless, these methods do not consider the presence of missing peaks and the bias produced by spectral overlapping.

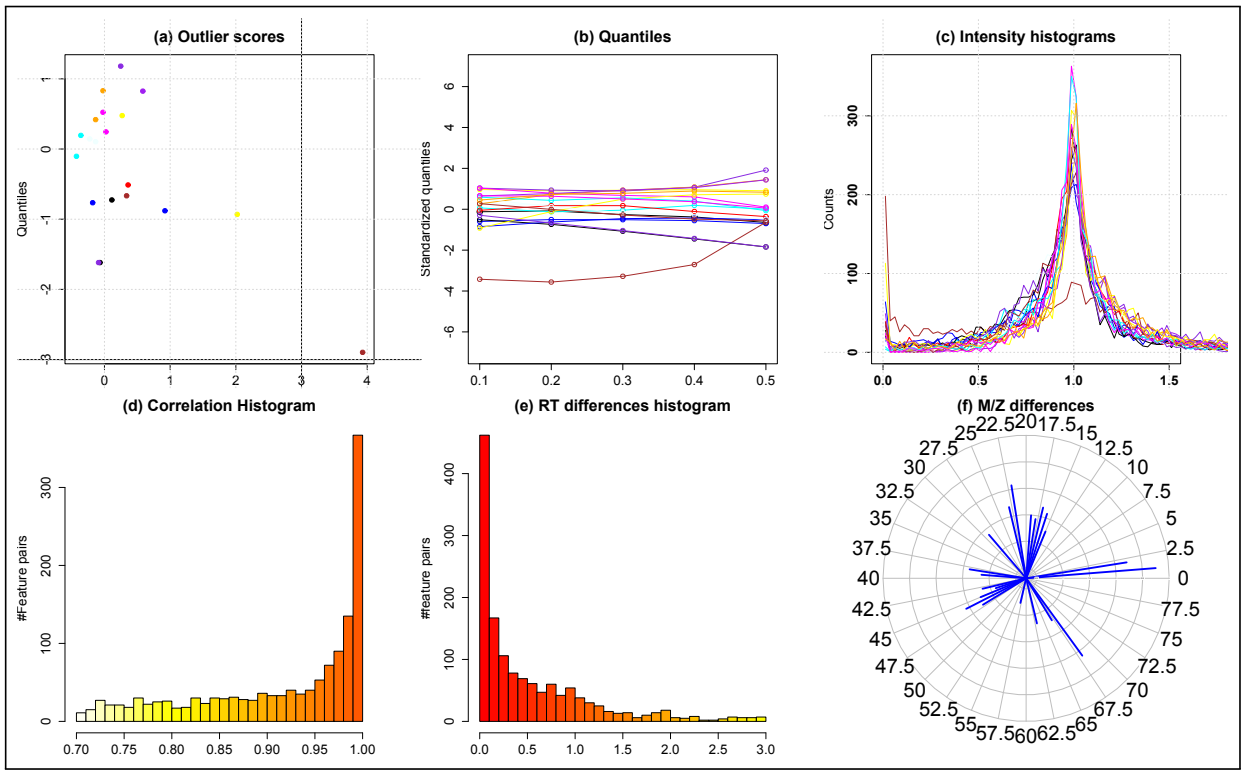


Figure 1.26: *AStream output.* This figure shows the first output of AStream method when analyzing a LC-MS dataset. Panels A to C show the results of the subsequent QC steps performed by AStream. After QC, AStream computes the intensity correlation between each pair of peaks. The resulting distribution shows enrichment for high correlation values that correspond to the presence of peaks coming from the same compound (D). These peaks are also characterized by low retention time differences (E). Finally, AStream identifies that peak pairs that correspond to known isotopes or adducts (F).

Metabolite spectral databases are essential for metabolite identification. The quality of the stored data as well as the number of metabolite spectra available in these databases is critical for the performance of identification algorithms. During the last years, multiple databases have been developed and the number of available metabolite reference spectra is continuously growing[137;138]. The Human Metabolome Database (HMDB) is perhaps the most extensive public metabolomic spectral database to date[139]. The HMDB stores >40,000 different metabolite entries, with exhaustive biological metadata and MS/NMR spectral references. In addition to spectral databases, several studies have also contributed to characterize the metabolome of multiple types of samples. Many of these reference studies are also exceptional resources of high quality data associated with the biofluid, tissue, or cell type of interest[105;140;141].

2 | Objectives of the PhD Thesis

The present PhD thesis had two main objectives. The first objective was to improve the bioinformatics tools available for processing and analyzing data from high-throughput genomics and metabolomics studies. Like most omics sciences, genomics and metabolomics are rapidly evolving research areas that constantly require new methods that are able to accurately extract the maximum amount of biological information out of very large and complex datasets. The main aim of this PhD thesis was to develop new analytical frameworks for these two types of omics datasets, providing advances in data integration, analytical accuracy as well as including new useful functionalities. Each methodological improvement included in this work started with a comprehensive and exhaustive evaluation of the available bioinformatics tools in order to identify their limitations. Simultaneously, the unmet needs of the researchers using these high-throughput technologies were also identified and taken into account when planning the development of the processing pipelines.

Achieving the objectives proposed in this PhD thesis has allowed the Rheumatology Research Group from the Vall d'Hebron Hospital to improve the analytical power of the genomic and metabolomic studies performed within the IMID Consortium. The IMID Consortium is a network of more than 80 clinical departments and biomedical research groups specialized in the characterization of the molecular basis of Immune-Mediated Inflammatory Diseases (IMIDs) through the use of high-throughput molecular data analysis ("omics") approaches. This biomedical consortium has been funded by the Spanish Ministry of Economy and Competitiveness grants (*PSE-010000-2006-6* and *IPT-010000-2010-36*). In these two projects, different high-throughput molecular data has been generated on large cohorts of phenotypically well-defined IMID patients and controls. Extensive genomic (>8,000 individuals) and metabolomic (>3,000 individuals) data has been generated using DNA genotyping microarray technology and proton nuclear magnetic resonance, respectively. Importantly, the new methodologies developed in this PhD thesis have been made publicly accessible as open-source software tools and, therefore, the entire scientific community will be also able to benefit from these technical advances.

The second main objective of this PhD thesis was to conduct a genome-wide association analysis (GWAS) of Crohn's disease (CD) phenotypes. CD is the most prevalent inflammatory

bowel disease and it is characterized by a marked heterogeneity, with patients presenting mild symptoms to patients showing a very severe affection. To date, very little is known about the genetic basis of phenotype heterogeneity in CD. The identification of genomic regions associated to CD clinical phenotypes is of high relevance, since it will allow a better understanding of the biologic mechanisms relevant for this clinical heterogeneity and it could lead to improvements in treatment personalization and the identification of more effective drug targets. Using the genotypic and phenotypic data of the patients included in the *IMID-Kit* project, the genetic risk for each phenotype has been studied, for the first time, using a genome-wide approach. The student has applied all the experience gained through the development of bioinformatics tools for processing genomics data to conduct the GWAS analysis. The interaction with the clinical specialists as well as the bioinformatics team of the *IMID-Kit* project has also been crucial component of the success of this study. The results provided by this study explain how genetic variation is associated with disease heterogeneity in CD, and provide an invaluable resource for the identification of the physiopathological mechanisms underlying each phenotype.

A full description of the objectives in each thesis section is presented below.

i **Development of a pipeline for SNP and CNV genotyping of Illumina microarray data for genome-wide association studies**

The main objective of this section is to develop a new SNP and CNV genotyping tool that overcomes the shortcomings of the previous methodologies. The genotyping pipeline will be based on the Illumina microarrays which is the most commonly used GWAS genotyping technology. The new algorithms implemented in this tool must provide higher genotyping call rate and accuracy in order to increase the power of high-throughput studies. Most CNV algorithms require multiple consecutive probes in order to provide accurate calls. The implemented methodology should be sensitive enough to perform CNV genotyping at the single probe level, and therefore be able to identify risk CNVs of smaller size.

The main objectives ofthe algorithms implemented in the new genotyping pipeline were:

- To develop a **new SNP genotyping method** that improves the SNP genotyping performance both in terms of genotyping call rate and genotyping accuracy.
- To develop a **new CNV genotyping method** that increases the CNV identification sensitivity and improves CNV genotyping call rate and accuracy.
- To perform an **exhaustive algorithm benchmarking** to compare the performance of the new SNP and CNV genotyping methods against state-of-the-art methodologies.

- To **integrate SNP and CNV genotyping** in a single processing pipeline. The resulting software tool must be **user-friendly and computationally efficient** in order to genotype hundreds or thousands of samples in a reduced amount of time.

ii **Development of a workflow for 1D-NMR spectra processing for high-throughput metabolomics studies**

The analysis high-throughput metabolomics studies based on NMR spectra represents a major processing challenge due to the lack of integrated and automatic processing tools. Although some methods cover a part of the processing workflow, some critical steps are still performed manually due to the low accuracy of the available automatic methods. Consequently, current high-throughput metabolomics studies are limited by the time required to process each sample, making it almost impossible to analyze large number of samples. The main objective of this part of the thesis was to develop a complete data processing pipeline for 1D-NMR metabolomic studies, and to improve several of the processing algorithms.

The specific objectives of the algorithms implemented in this processing pipeline are:

- To **integrate the processing stages** that are required in a typical metabolomics NMR analysis. This automated workflow will include pre-processing, quality control, spectral alignment, peak detection, peak quantification, and metabolite identification.
- To provide a **computational efficient tool** that will be able to perform high-throughput analyses (>100-1000 samples).
- To develop a **spectral alignment algorithm** that can deal with the large inter-sample variance present in 1D-NMR data. The alignment algorithm will avoid the use of a reference spectrum, which often leads to a reduction in the alignment performance.
- To develop a **metabolite identification algorithm** that uses all the reference spectral data available in publicly available databases. This reference data will be used to match the peak patterns of each metabolite within the analyzed spectra and facilitate the identification of the metabolites of interest.
- To provide a **final quantification matrix** ready for downstream metabolomics statistical analysis.

iii **Genome-Wide Association Analysis of Crohn's Disease Phenotypes**

The identification of the biological basis of Crohn's disease heterogeneity could have a major impact in the development of new clinical and pharmacological approaches

that can more efficiently control disease severity and, therefore, improve the quality of life of patients with this inflammatory bowel disease. Although previous studies have analyzed the association of known CD risk loci with the risk to develop different clinical phenotypes, to date, this has not yet been investigated at a genome-wide scale. The objective of this study was therefore to perform the first genome-wide association study (GWAS) to identify risk loci for Crohn's disease phenotypes. The identification of these risk loci will demonstrate the existence of a genetic component that is specific for disease heterogeneity. The specific objectives for this study are:

- To define the **phenotypes that are of high clinical relevance** in Crohn's Disease. To accurately define these phenotypes, the collaboration with clinical specialists will be essential.
- To curate the **clinical database** and extract the phenotypes for the patients included in the study.
- Perform the first **GWAS analysis** (i.e. 576,818 SNPs) for each phenotype in a well-characterized discovery cohort of 1,090 Crohn's disease patients from Spain.
- Select the **candidate risk loci** to be replicated in an independent validation cohort of 1,296 patients from Spain.
- To identify the **biological mechanisms** that could explain the observed genetic association with CD clinical phenotypes.

3 | GStream: Improving SNP and CNV Coverage on Genome Wide Association Studies

Note: This chapter is an exact copy of the paper:
A. Alonso, S. Marsal, R. Tortosa, O. Canela-Xandri and A. Julià. **GStream: Improving SNP and CNV Coverage on Genome-Wide Association Studies.**
PLoS ONE, 2013, 8(7), e68822. DOI: 10.1371/journal.pone.0068822.

Abstract

We present GStream, a method that combines genome-wide SNP and CNV genotyping in the Illumina microarray platform with unprecedented accuracy. This new method outperforms previous well-established SNP genotyping software. More importantly, the CNV calling algorithm of GStream dramatically improves the results obtained by previous state-of-the-art methods and yields an accuracy that is close to that obtained by purely CNV-oriented technologies like Comparative Genomic Hybridization (CGH). We demonstrate the superior performance of GStream using microarray data generated from HapMap samples. Using the reference CNV calls generated by the 1000 Genomes Project (1KGP) and well-known studies on whole genome CNV characterization based either on CGH or genotyping microarray technologies, we show that GStream can increase the number of reliably detected variants up to 25% compared to previously developed methods. Furthermore, the increased genome coverage provided by GStream allows the discovery of CNVs in close linkage disequilibrium with SNPs, previously associated with disease risk in published Genome-Wide Association Studies (GWAS). These results could provide important insights into the biological mechanism underlying the detected disease risk association. With GStream, large-scale GWAS will not only benefit from the combined genotyping of SNPs and CNVs at an unprecedented accuracy, but will also take advantage of the computational efficiency of the method.

3.1 Introduction

Over the last years, Genome-Wide Association Studies (GWAS) using microarray-based technology have played an important role in the identification of common genetic variations and their relationship with disease susceptibility[18–21]. Genotyping microarrays[22] have enabled this success through the parallel genotyping of thousands of Single Nucleotide Polymorphisms (SNPs), capturing most of the common variation in the human genome. Very recently, the new generation of microarrays have integrated the extensive knowledge revealed by the 1KGP[53] and, together with the decreasing costs of this technology, are now allowing the use of the GWAS approach to the association of rare genetic risk variants or more complex human traits.

In addition to SNPs, Copy Number Variants (CNVs) have shown to play an important role in disease susceptibility[26]. CNVs are relatively large (> 500 bp) genomic variations and include deletions, tandem duplications and insertions[42]. Recent studies based either on specific CGH arrays or genotyping microarrays have demonstrated the importance of CNVs due to their global contribution to the human genome variation, their functional impact and their role in human disease[24;26–30;142]. Some of these reference studies have contributed to elaborate a map of regions containing highly polymorphic CNVs called Copy Number Polymorphisms (CNPs)[24;142;143]. These common variations have appeared as a significant area of interest, since they segregate in the population at an appreciable frequency and their analysis over big sample collections could potentially lead to significant disease risk associations.

The development of the two mentioned technologies (CGH arrays and genotyping microarrays) for high throughput CNV screening has prompted the inclusion of CNVs in GWAS studies[31;144–146]. When comparing both technologies, genotyping microarrays offer the practical advantage of obtaining at the same time SNP and CNV genotype data. However, there is still a major need to develop methods that can best deal with the signal-to-noise ratio deficiencies and genomic coverage of genotyping microarray data when attempting to identify and quantify CNVs. So far, most of the commonly used methods for CNV detection and genotyping at the genome-wide scale[76;77;147] are based on independent per-sample analyses that use summarized measurements relative to a reference set of samples. This type of approach has proven to work well for large genomic variations, but it fails to use the powerful information generated by analyzing multiple samples, leading to high false negative rates with small CNVs[78].

In this study we present GStream, a method for SNP and CNV/CNP genotyping that is tailored to GWAS objectives. GStream integrates a substantially improved version of our

previous CNV calling software CNstream[148]. Our new method achieves a superior accuracy in both SNP and CNV genotyping compared to well-established methods. Indeed, GStream obtains an unprecedented accuracy within CNV regions, with a performance close to that obtained from purely CNV-oriented technologies like CGH arrays. All these improvements have been quantitatively compared against previous state-of-the-art methods and accurately assessed using different Illumina genotyping microarrays together with publicly available SNP[51] and CNV[24;142;143] reference datasets based on Next-Generation Sequencing (NGS), CGH array and genotyping microarray technologies. Finally, the computational efficiency of the method has been optimized, enabling the large-scale SNP and CNV analyses to be performed in a short amount of time.

In addition to presenting this new method and demonstrating its superior performance over reference datasets, we have also performed different relational analyses concerning previously known risk loci. Using GStream we have been able to identify, for the first time, several CNVs in strong linkage disequilibrium (LD) with risk-associated SNPs[149] as well as CNVs spanning disease-associated genes[150]. Together, these results could reveal important insights into the causality of these disease risk associations.

3.2 Material and methods

We first introduce the Illumina BeadChip microarrays and describe the algorithms for SNP and CNV genotyping. Next, we provide information about the datasets used in this study and the implemented metrics for evaluating SNP and CNV genotyping accuracy. Finally, we describe the methods used for the CNV association studies that we have run using the GWAS catalog[149] and the OMIM[150] databases.

3.2.1 Illumina BeadChip Data

Illumina BeadChips have been largely used in large-scale genome-wide association studies and are based on the Infinium assay technology[56]. This type of genotyping array consists on hundreds of thousands of probe pairs designed to capture genomic variation at the SNP and CNV level. In each probe pair, each probe has been designed to specifically bind one of the two SNP alleles (i.e. alleles A and B) generating a pair of fluorescence intensities. These signals are then measured and processed in order to infer the presence or absence of these alleles in each sample. GStream software uses these raw measurements to extract SNP and CNV genotypes for each sample at each probe pair. From here on, fluorescence measurements of alleles A and B will be called channel A and B intensities and samples will be categorized at each SNP as homozygotes (i.e. AA or BB) or heterozygotes (i.e. AB).

3.2.2 GStream method for SNP genotyping

Before identifying the clusters corresponding to the AA, AB and BB genotypes at each probe, raw intensity data of each probe must be normalized in order to equalize the overall sample intensity distribution at each channel (Figure 3.1). This step is crucial since the sensitivity differences of each probe and channel can lead to bias affecting the genotyping performance. The method used by GStream is based on the scaling correction used by Peiffer et al.[75]. In this method, the intensity centroids of a set of pre-computed AA and BB candidate homozygote samples are identified and used to scale channel A and B intensities. However, GStream adds two modifications in order to improve the normalization in those cases involving probes capturing both SNP and CNV variation. First, instead of using candidate homozygote intensity centroids, the scaling parameter is computed by weighting the candidate homozygotes intensity distributions (the higher the intensity, the higher the weight) and by finding the maximum over the resulting curve (Supplementary Figure A.1). This modification helps GStream to better deal with the particularities of intensity distributions coming from probes within CNV regions. The second modification introduced by GStream uses heterozygote intensity data when no homozygote candidates are found, thus helping to better deal with probes capturing low MAF SNPs.

Once the intensities from both A and B channels have been normalized, GStream proceeds to identify the clusters corresponding to each SNP genotype (i.e. AA, AB and BB). Developing an accurate SNP genotyping method is crucial not only for SNP analysis itself, but also because GStream CNV genotyping method uses this information to improve the accuracy of the CNV call. GStream applies the following procedure to assign a SNP genotype to each sample at each marker:

1. Channel A and channel B intensities from the analyzed marker are transformed to B allele frequencies (BAF) and absolute intensities[75].

2. Absolute intensities are used to detect samples without any allele (homozygous deletions) which are characterized by very low intensities at both channels. In order to compute the absolute intensity threshold between homozygous deletion samples and the other samples, the absolute intensities are sorted and then differences between each pair of consecutive intensities are computed. If high intensity differences are found within the expected threshold range ($[0, 0.5]$), the zero-threshold is fixed to the corresponding intensities (Supplementary Figure A.2).

3. The BAF probability density function (PDF) is estimated by computing the scaled histogram of all the sample BAF values. Peaks over this PDF corresponding to genotype

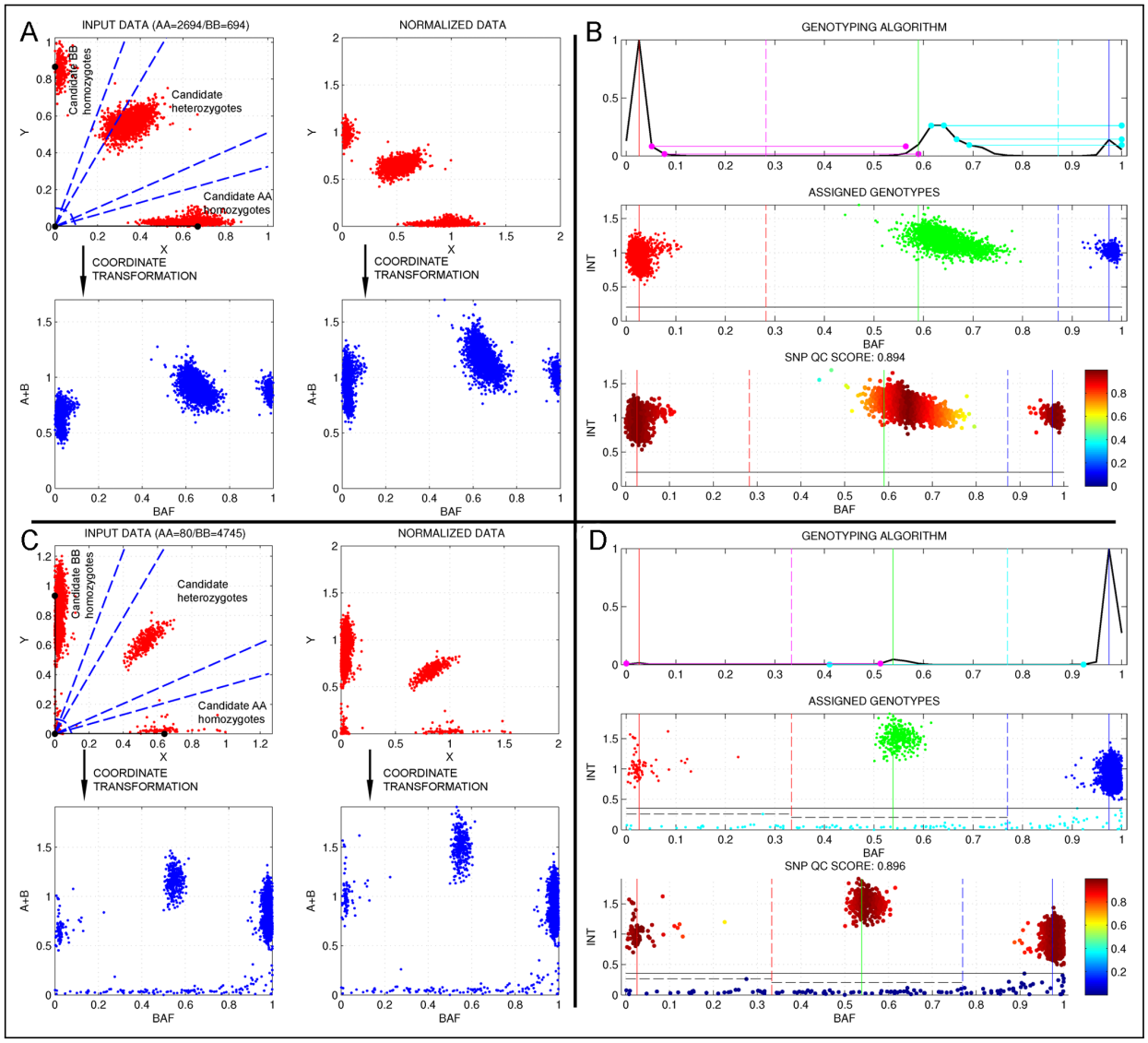


Figure 3.1: *GStream method for SNP genotyping.* This figure shows how GStream genotyping method works on two example markers, the first one representing a typical marker capturing a SNP (A and B) and the second one capturing both a SNP and a CNV (C and D). The leftmost graphs show the effects of the normalization procedure for the two markers, where the dotted blue lines enclose the ranges where candidate homozygotes and heterozygotes are identified in order to compute the scaling factors for each channel (black points over the axes). The rightmost graphs give an overview of the genotyping procedure: Upper subfigures represent the scaled BAF probability density function with the solid vertical lines setting the identified genotype centres, the dotted vertical lines setting the genotype limits and the horizontal lines representing the sequential search of genotype cluster peaks. Medium and lower subfigures represent genotype calls and quality call scores respectively.

clusters will be identified sequentially starting by the peak generated by the major allele frequency homozygote cluster. The algorithm establishes a minimum separation between peaks in order to assign them to different genotype clusters and it stops when three peaks have been found or when no more peaks are found. Once genotype peaks have been found, genotype limits are computed by finding the PDF minimum

between each consecutive pair of centres (Figure 3.1). These limits will define the BAF intervals assigned to each genotype and each sample will be genotyped accordingly to them.

4. If the number of genotype peaks identified is less than three, each genotype cluster is re-analyzed with a better resolution (i.e. increasing the number of histogram bins to estimate the BAF PDF) in order to identify sub-clusters corresponding to different genotypes. This procedure avoids common errors seen in others algorithms where, for example, genotypes of SNPs with highly discordant sensitivities at each channel are incorrectly assigned.

5. Finally, a global genotyping quality score and an individual score for each sample genotype are computed (Figure 1). The global score is proportional to the standard deviation mean of the BAF values assigned to each genotype and the individual score corresponds to the distance between the sample BAF value and its corresponding genotype peak divided by the distance between genotype centres.

Both genotypes and quality control measurements for each sample are extracted by GStream to the output files. This information is also required by the CNV genotyping method, which is based both on the normalized channel intensities and the SNP genotype information. Further algorithm details are given in Supplementary Text A.3.

3.2.3 GStream method for CNV genotyping

CNV identification and genotyping is one of the principal contributions of GStream to the current state-of-the-art microarray genotyping methodology, clearly outperforming previous approaches. Although this method has been based on our previous CNstream method[148], multiple changes have been introduced in order to improve performance and computational efficiency.

GStream uses normalized intensities and SNP genotypes computed in the SNP genotyping stage to identify the presence of deletions and amplifications characterized by variable clustering patterns on the intensity data (i.e. high frequency CNVs) or by slight deviations from the diploid distribution (i.e. low frequency CNVs).

One of the improvements incorporated in the algorithm is that each SNP genotype cluster is independently analyzed taking only into account the intensity channel that carries valuable information. This way, the CNV algorithm is divided in four parallel steps (Figure 3.2A):

1. Analysis of channel A intensities from the samples previously genotyped as AA homozygotes.

2. Analysis of channel B intensities from the samples previously genotyped as BB homozygotes.

3. Analysis of channel A intensities from the samples previously genotyped as heterozygotes (i.e. AB).

4. Analysis of channel B intensities from the samples previously genotyped as heterozygotes (i.e. AB).

As well as dividing the analysis in four independent steps, the algorithm is based on the following assumptions:

1. Homozygous deletions (0 allele copies) are previously detected during the SNP genotyping stage.

2. Due to the technical limitations of genotyping microarrays, the intensity measurements show a saturation effect when amplifications are found. For this reason, intensity clustering patterns corresponding to amplifications are very rare and hard to detect unless they span multiple probes[151].

3. Samples categorized as homozygote samples (i.e. AA and BB) can correspond to hemizygous deletions (i.e. A and B) or amplifications (i.e. AA+ and BB+). Due to the saturation effect the algorithm does not stratify amplifications by the number of allele copies.

4. Samples characterized as heterozygotes (i.e. AB) can have two or more copies (i.e. AB, AAB, ABB...). The total number of copies can be inferred by independently computing the number of copies of each allele and then adding the results for each sample.

Below we describe the procedure for determining the CNV genotypes from the set of channel intensities of each one of the four analysis steps.

Model selection

For each SNP genotype, the algorithm starts identifying clusters over the channel intensities that carry the corresponding allele information (i.e. channel A for AA homozygotes, channel B for BB homozygotes and both channels for AB heterozygotes). Due to the mentioned saturation effects, it is very uncommon to observe more than two intensity clusters in microarray data and, for this reason, only two models will be fitted to the intensity data: a one- and a two-component Gaussian mixture model (GMM). The first one will be fitted

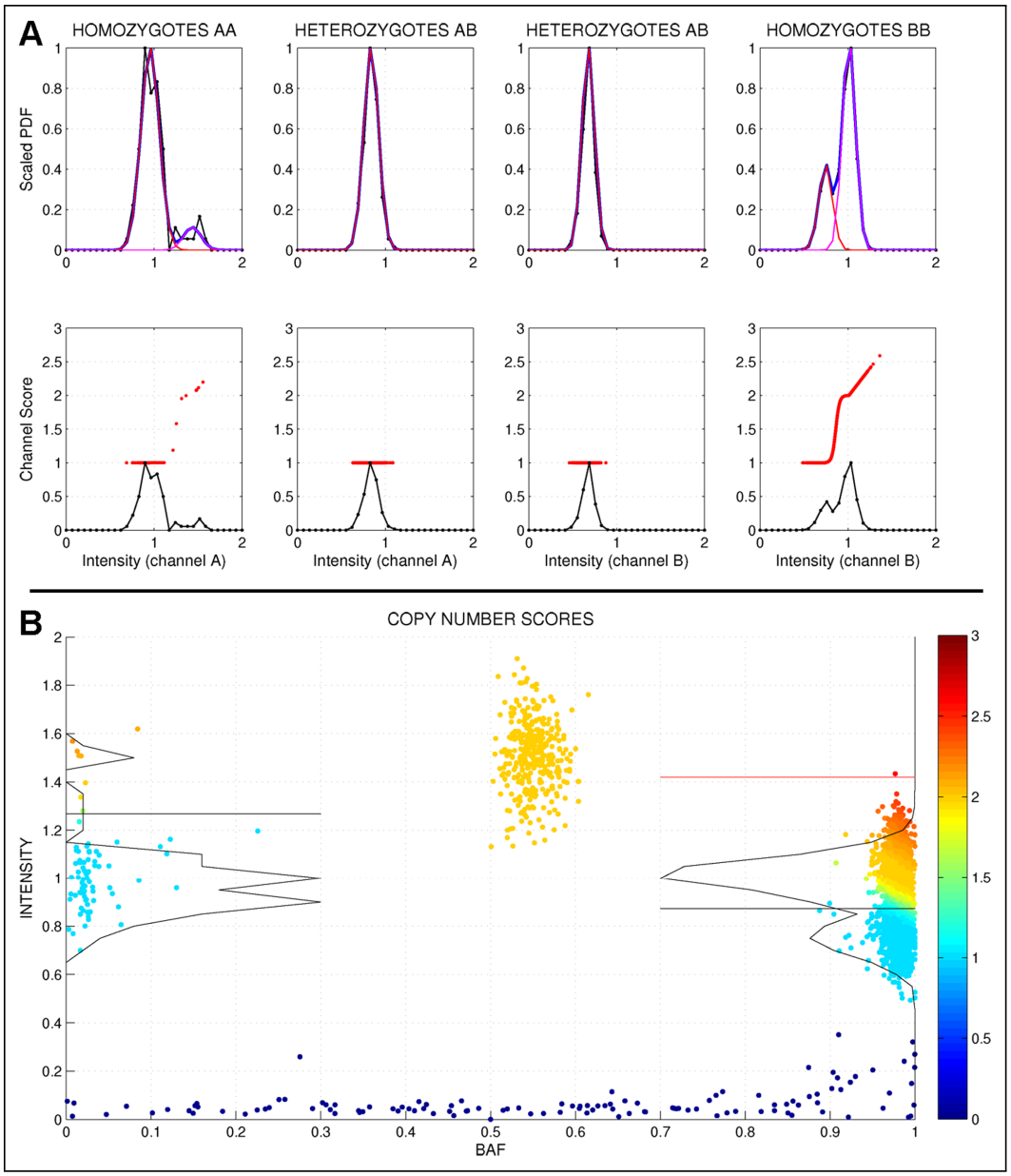


Figure 3.2: *GStream method for CNV genotyping.* (A) Each CNV analysis is divided in four independent sets where the number of allele copies per channel intensity is estimated. Here, the homozygote intensities over its respective informative channels (upper rightmost and leftmost graphs) are fitted with a two-component model (in this case, capturing a deletion) while heterozygote intensities over each channel are better fitted with a one-component model (upper centre graphs). Lower graphs show the intensity distributions (solid black lines) together with the corresponding copy number score (red points) assigned to each sample. AA homozygotes are mostly classified as deletions (scores near to 1), BB homozygotes are divided into diploids (scores∼2) and deletions (scores∼1) while heterozygotes are classified as diploids (i.e. one allele detected at each channel). (B) Final representation of the analyzed probe where points represent samples and colour their relative copy number scores. SNP and CNV genotypes are assigned along the BAF and the intensity axis respectively.

using the mean and the variance of all the intensities while the second one will be fitted using the Expectation-Maximization algorithm[152] (Figure 3.2A). A set of requirements in

order to select the second model have been carefully developed and only if all of them are accomplished, the two-component model (indicating a pattern corresponding to a common CNV) will be selected (Figure 3.2A).

Component labelling

If the two-component model has been selected, a copy number category will be assigned to each one of the two components. As no prior knowledge is available to assign the two components either to a deletion pattern (i.e. $CN = 1$ and $CN = 2$) or to an amplification pattern (i.e. $CN = 2$ and $CN = 3$), a disambiguation method is necessary. GStream bases the component labelling both on the relative weight of each component (i.e. proportional to the copy number frequency) and on the presence of homozygous deletions (Supplementary Figure A.3A). When the one-component model has been selected, the component will be labelled by default to $CN = 2$ (i.e. diploid), which is assumed to be the most common state.

Outlier identification and CNV scoring

Outlier identification is intended to capture low frequency CNVs that are not captured by a two-component GMM and is based on identifying samples showing high or low deviations from the intensity distributions defined by the selected model. CNV scoring assigns to each sample i a score S between 0 and 3 depending on its copy number posterior probabilities (Supplementary Figure A.3B). At the end of this step, GStream has obtained a CNV score for all the samples that allows the identification of deletions and amplifications as well as a quantification of the reliability of the assignment (Figure 3.2B). Additional algorithm details are given in Supplementary Text A.3.

3.2.4 Microarray data from HapMap samples

In order to evaluate and compare the performance of GStream SNP and CNV genotyping methods we have used raw Illumina microarray data from HapMap samples available at the Gene Expression Omnibus (GEO, *http://www.ncbi.nlm.nih.gov/geo/*) [153] (Table 3.1). This data has been also used as the input of state-of-the-art SNP and CNV genotyping algorithms in order to extract accurate comparison measures as well as to ensure a performance assessment independent from the technical biases of the raw data. Only markers having available NCBI Build 37 mapping information were kept for further analysis.

Platform	Samples[1]	GEO accession[2]	Autosomal markers	Evaluation
Human610-Quad	73/75/77	GSE17205/GSE17206/GSE17207	596528	568182
Human660W-Quad	89/89/89	GSE17208/GSE17209/GSE17210	638582	552529
Human1M-Duo	89/90/90	GSE16894/GSE16895/GSE16896	1141594	1058827
HumanOmni1-Quad	88/89/90	GSE17197/GSE17201/GSE17203	1103791	882445

[1] Number of samples in the CEU, CHBJPT and YRI population datasets.
[2] GEO accession numbers of the CEU, CHBJPT and YRI population datasets.

Table 3.1: Public microarray data used in this study. The used microarray data comes from four different Illumina BeadChip platforms and the sample data comes from three HapMap populations. The total number of autosomal markers and the number of markers used for SNP genotyping evaluation are shown.

3.2.5 SNP genotyping performance evaluation and comparison with previous methods

Golden standard genotypes

In order to correctly evaluate the SNP genotyping algorithm performance, a set of independent and high-quality genotype calls is required. The genotype calls of HapMap samples have been established as a golden standard commonly used in the literature for performance evaluation of SNP genotyping methods. These calls are available for download through the online HapMap tool HapMart (*http://hapmap.ncbi.nlm.nih.gov/biomart*). For this study, we downloaded the genotypes corresponding to the samples having available microarray data and used them as the golden standard. SNPs used for performance evaluation were chosen in order to fulfil three criteria: (i) to have available Build37 mapping information, (ii) to be present both in the analyzed microarray platform and in the golden standard HapMap dataset, and (iii) to have concordant reference alleles both in the microarray and in the golden standard annotations (Table 3.1).

Algorithms

GStream SNP genotyping accuracy has been evaluated and compared with three methods: (i) GenoSNP[73] which is a well-known genotyping algorithm based on a within-sample approach; (ii) GenCall, which is the proprietary (Illumina, San Diego, US) algorithm[72], and it is used by the vendor genotyping software; (iii) M3[74], which is a recently published method for SNP genotyping that re-analyzes the data in order to increase the accuracy over the low MAF SNPs and has shown to have increased performance.

3.2.6 Copy number genotyping performance evaluation

Copy number evaluation was performed at two levels: Evaluation of the GStream ability to detect structural variation obtained from the 1000 Genomes Project[53] next-generation sequencing (NGS) data and evaluation and comparison of CNV population association results using different algorithms and golden standard calls from three recently published studies[24;142;143]. Below we describe materials, methods and metrics used for this two-stage evaluation.

Evaluation of CNV genotyping accuracy over the 1KGP Structural Variants

In order to test the ability of GStream to detect copy number variation, we have used the HumanOmni1-Quad GStream calls and a golden standard dataset from a public release of the 1KGP. HumanOmni1-Quad platform was chosen due to its highest coverage and resolution which allowed an evaluation over a major number of loci. The golden standard dataset consisted of the last variant call files that have been released by the 1KGP (version v3/20110521, *ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release*). From all the variants included in the 1KGP release we chose those that corresponded to structural variations (i.e. CNVs) and we filtered out variants with a MAF lower than 2% within all the populations. The resulting set included 2,531 structural variants (SV) with their respective calls over N=353 unrelated HapMap samples. From these 2,531 loci, 1,956 are covered by HumanOmni1-Quad markers and GStream calls were available for 149 out of the 353 1KGP samples, which jointly formed the final evaluation set.

The evaluation procedure consisted of finding the markers whose GStream calls are in maximum LD with each SV:

$$\max_{i \in S_k} \{r^2(CN_i, SV_k)\} \tag{3.1}$$

where S_k is the set of microarray markers within the region spanned by the SV, CN_i is the copy number genotypes assigned by GStream at marker i and SV_k is the 1KGP calls for the analyzed SV k.

Evaluation of the power to detect genome-wide associations over CNV markers and comparison with previous methods

The objective of this section is to evaluate the power to detect CNV associations and to compare GStream performance with other well-known methods. This comparison was performed using Human1M-Duo and HumanOmni1-Quad platforms. Using these two platforms also allowed an assessment of the specific platform power to detect CNV associations,

comparing platforms with (i.e. HumanOmni1) and without (i.e. Human1M) a specific design to cover CNV.

The genome-wide association study over CNV markers was performed at the population level aiming to identify CNVs significantly associated to specific populations and comparing the association statistics with those obtained from golden standard datasets.

The CNV algorithms used are described below:

- PennCNV[76] is one of the most frequently used methods for analyzing CNVs using Illumina microarrays. This software implements a CNV estimation method based on Hidden-Markov-Models (HMMs), in which copy number calls are performed sample by sample by analyzing the sample LRR (i.e. absolute intensity) and BAF values at each marker. Default settings were used in the analysis of the available HapMap samples generating the PFB file (i.e. population frequency of B allele) from their genotyping data.
- QuantiSNP is also one of the well-known methods for CNV analysis over Illumina microarrays. It is based on an Objective Bayes Hidden-Markov Model that is used to set certain hyperparameters in the HMM priors (for details see Colella et al.[77]). Default settings were used in this analysis with the provided Infinum HD parameter files and the local GC content files.
- CNstream[148] was also evaluated in order to demonstrate how our new method overperforms the previous one due to the major critical modifications introduced.

Association statistics obtained by each algorithm were compared with those obtained from three recently published reference studies:

- The first dataset was obtained from a study published by McCarroll et al.[143]. In this study a hybrid genotyping array was designed to simultaneously measure SNPs and CNVs. Almost half (N=1,320) of the targeted CNV regions were observed in multiple unrelated individuals and were defined as CNPs. From this set of CNPs we selected the autosomal CNPs (N=1,292) over the 270 HapMap samples as the first golden standard dataset.

- We used the data published by Conrad et al.[24] as the second golden standard dataset. In this study, an Agilent CGH based array was used to generate a map of CNVs greater than 443 base pairs. For 4,978 of these CNVs reference genotypes from 450 HapMap samples are available to download. We used the corresponding sample calls for all the 4,899 autosomal CNVs.

- The last dataset used for CNV performance evaluation was obtained from the results published by Campbell et al.[142]. In this study a custom Agilent CGH microarray targeting regions of known CNPs was designed and evaluated over HapMap samples of diverse ethnic backgrounds. For this analysis we used the published discrete CNV calls of polymorphic loci represented in the reference genome assembly (N=874) for the 487 HapMap samples included in the analysis.

In order to provide a measure of genome-wide association power, pairwise population-association tests (CEU:YRI, CEU:CHBJPT and YRI:CHBJPT) were performed using the calls from the three golden standard datasets. Loci that either were not covered by the microarray platform or did not obtain significant associations (*P*-Value < 0.05) in any population test (Table 3.2) were filtered out. Chi-square test *P*-Values were then computed at each locus and compared to those obtained using the calls of the four methods across the markers covering the locus. For each algorithm, the marker obtaining the best result across the region was selected for comparison.

Study	Technology	N_{CNVR} [1]	HumanOmni1-Quad		Human1M-Duo	
			Coverage	N_{ASSOC} [2]	Coverage	N_{ASSOC} [2]
McCarroll	Affymetrix	1292	1290	929	1288	927
Campbell	Agilent CGH	874	874	962	873	962
Conrad	Agilent CGH	4899	4899	3659	4899	3671

[1] N_{CNVR} refers to the number of CNV loci selected from each study.
[2] N_{ASSOC} column refers to the total number of associated regions (*P*-Value<0.05) for the three population tests detected over the golden standard calls.

Table 3.2: CNV regions for each dataset and platform used to evaluate the power to detect genome-wide associations. Coverage with at least one marker within the CNV loci of both platforms is very similar although the marker density differs considerably.

3.2.7 Copy number variation and disease susceptibility

Using microarray data to accurately extract information from copy number variation can be particularly relevant when trying to identify all the type of variants that convey risk to disease susceptibility. Using two available catalogues of disease genomic associations we have used two different approaches to demonstrate the joint capacity of microarray platforms and the GStream method to identify new CNV disease associations. The two analysis explained here have been performed using the CNV calls inferred by GStream over the HapMap samples genotyped with the Illumina HumanOmni1-Quad platform.

Catalog of published genome-wide association studies

Since microarray genotyping platforms became available, a large number GWAS have allowed the discovery of important SNP-trait associations. However, some of these SNPs have limited or no known functional impact. In these cases, the possibility that they act as proxies of other types of variations (i.e. CNVs[154;155]) with a deeper functional impact is more likely.

In order to identify putative causal CNVs we have analyzed the LD patterns between all the trait-associated SNPs reported by the catalog of published genome-wide association studies (*http://www.genome.gov/gwastudies*)[149] and the CNV microarray markers detected over the HumanOmni1-Quad platform. Trait-associated SNP genotypes were extracted from the 1KGP data reported previously and CNV genotypes were called with GStream. All the HumanOmni1-Quad markers that presented a non-diploid frequency greater than 1% (CNV markers) were included in the analysis (NCNV=90,892) together with the 7,571 trait-associated SNPs.

The conditions used for selecting the candidate SNP-CNV pairs where the CNV could provide new functional information on the reported association are described below:

1. Distance between the SNP and the CNV markers not greater than 50kb.
2. Correlation coefficient r^2 greater than 0.7 in any of the three analyzed HapMap populations (CEU, YRI and CHB+JPT).
3. Distance between the CNV marker and the nearest gene not greater than 100kb or CNV marker spanning binding transcription factor regions as defined by the Transcription Factor ChIP-seq track on the UCSC browser[156].

From the 7,571 trait-associated SNPs, 382 were paired with one or more CNV markers fulfilling these conditions. A final set of 333 SNP-CNV pairs was obtained after filtering out repeated associations of SNPs with the same trait by different GWAS studies.

CNV overlapping analysis with disease-related genes

In this second approach, we examined the CNV variants called by GStream spanning genes known to be involved in disease based on the OMIM database (*http://www.omim.org*)[150]. In order to characterize CNVs with a high probability of conveying functional effects on the disease-related OMIM genes we set multiple strict selection criteria:

1. From the initial set of CNV markers ($N_{CNV} = 90{,}892$) only those located less than 15kb away from an OMIM gene and with at least two more CNV markers covering this gene were selected ($N_{CNV} = 5{,}836$).

2. We defined CNV loci as sets of three or more nearer CNV markers (i.e. distance between them not greater than 5kb) in high LD ($r^2 > 0.7$) spanning the same OMIM gene. After applying this filter we obtained a final set of 212 CNV loci spanning OMIM genes.

3. Finally, when more than one CNV locus spanned the same gene, only the one showing the greatest r^2 measurements between its CNV markers was kept for further analysis.

The final set of candidates consisted of 149 CNV loci spanning disease-related OMIM genes.

3.2.8 Software availability

An executable version of GStream along with the documentation and example data files can be freely downloaded from our website *http://www.urr.cat/GStream*. This web site also provides regularly updated results of new CNV associations within known human risk loci identified with this method.

3.3 Results

3.3.1 Performance assessment of SNP genotyping

For each available Illumina platform, the golden standard genotype calls were compared with the calls generated by GStream, GenoSNP, GenCall and M3 software tools. The global accuracy results over autosomal SNPs (Table 3.3) show a moderate improvement for GStream with respect to GenoSNP and a substantial improvement with regards to GenCall and M3 methods. GenCall performed very well when "non-called" genotypes where discarded, but its global performance decreased due to its low call rate. M3 algorithm could only be evaluated over the Human610-Quad and the Human660W platforms due to code incompatibilities with the Human1M-Duo and the HumanOmni1-Quad platforms. Although the improvement of GStream SNP-genotyping method regarding its competitors may not appear very high, they can represent a significant improvement from an absolute point of view (i.e. the accuracy differences when using HumanOmni1-Quad would be equivalent to a gain of 2,300 completely genotyped SNPs). Chromosome X genotyping accuracy was also evaluated, obtaining a similar decrease in performance ($\sim$0.5%) for all the algorithms and maintaining the accuracy differences between algorithms.

The second performance test consisted of computing the global accuracies at different levels of drop rate, where drop rate refers to the percentage of markers which are removed from the accuracy computation based on low call confidence measures (as defined by Ritchie et

Method	Call rate (%)[1]	Accuracy (%)[2]	Global accuracy (%)	CNstream differential (%)
Human610-Quad				
GStream	99.952	99.798	99.75	
GenoSNP	100	99.577	99.577	-0.173
GenCall	96.724	99.868	96.596	-3.154
M3	99.584	99.585	99.171	-0.579
Human660W-Quad				
GStream	99.99	99.804	99.794	
GenoSNP	100	99.65	99.65	-0.144
GenCall	95.411	99.879	95.295	-4.499
M3	99.798	99.635	99.434	-0.36
Human1M-Duo				
GStream	99.97	99.768	99.738	
GenoSNP	100	99.561	99.561	-0.177
GenCall	98.024	99.825	97.853	-1.885
M3	NA	NA	NA	NA
HumanOmni1-Quad				
GStream	99.971	99.671	99.643	
GenoSNP	100	99.435	99.435	-0.208
GenCall	97.083	99.747	96.838	-2.805
M3	NA	NA	NA	NA

[1] Percentage of called genotypes.
[2] Number of correct genotypes over the number of called genotypes.
[3] Number of correct genotypes over the total number of genotypes available within the golden standard dataset.

Table 3.3: Global accuracy results for SNP genotyping.

al.[157]). When markers are discarded by low global marker quality score -a common filtering procedure in GWAS quality control steps- GStream reaches the best performance for all the evaluated drop rates (Figure 3.3). Furthermore, the difference in accuracy between low and high drop rates is much lower on GStream, which implies a robust genotyping performance, even for those markers with lower quality scores. The low accuracy values obtained using GenCall at low drop rates can be explained from its low call rate (i.e. only when the drop rate exceeds the uncall rate, GenCall performance is comparable to the other algorithms, otherwise uncalled SNPs are included in the performance evaluation). When discarding markers by low quality sample calls, the results show a similar pattern but with reduced accuracy differences between the algorithms (Supplementary Figure A.4A). These results place GStream as the best option for SNP genotyping, since its genotyping accuracy reaches its maximum at lower drop rates compared to the other algorithms.

We also examined the performance with respect to the minor allele frequency (MAF, Figure S4B). Two key conclusions result from this analysis: First, probes capturing rare SNPs (MAF < 0.05) showed a slight accuracy reduction ($\sim$0.5%) on all the platforms and algorithms tested and, second, GStream accuracy gain with respect to the other algorithms was practically independent of the SNP MAF.

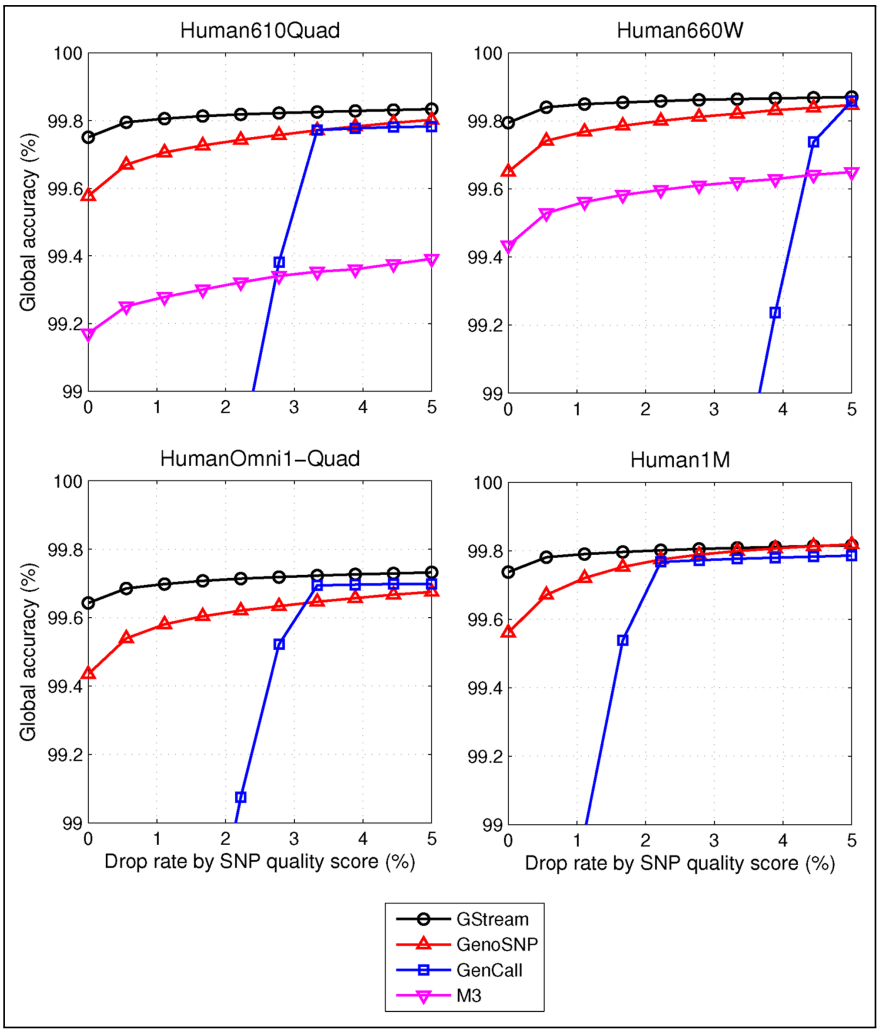


Figure 3.3: *Evaluating SNP genotyping performance.* Plots comparing SNP genotyping algorithms for each microarray platform are tested. The vertical axis represents the percentage of SNPs that are excluded from the accuracy calculation by the lowest quality score criteria. GStream performed better at all the drop rate levels in all the platforms. A high decrease in performance is observed for GenCall when drop rate values are lower than its uncall rate (i.e. ∼2% in Human610Quad).

Finally, we tested the effect of sample size on GStream accuracy (Supplementary Figure A.4C). The computed accuracy was compared to the accuracies obtained for the other algorithms when using all the Human610-Quad samples (N=225). However, even for sample sizes as low as N=20, the global accuracy of GStream is clearly higher than the accuracies obtained by the other algorithms, even if the highest sample size (N=225) is used, demonstrating the superior sensitivity of GStream genotyping algorithm.

3.3.2 Performance assessment of CNV genotyping

1KGP Structural Variants

SV calls from 1KGP for 149 unrelated HapMap samples (i.e. N_{CEU}=32, N_{YRI}=37 and N_{CHBJPT}=80) were compared with their respective GStream calls in order to measure the ability to detect

this type of variation using GStream on microarray data. For each SV (N=1,956), we computed the CNV genotyping accuracy by finding the maximum LD measurement between its golden standard calls and the GStream calls over the HumanOmni1-Quad markers covering the region. The results showed a high correlation between GStream and 1KGP calls: 75.7% of the SVs were captured by GStream with an r^2>0.8, 18.3% with an r^2<0.8 and only 6.0% were not detected by GStream (Figure 3.4A).

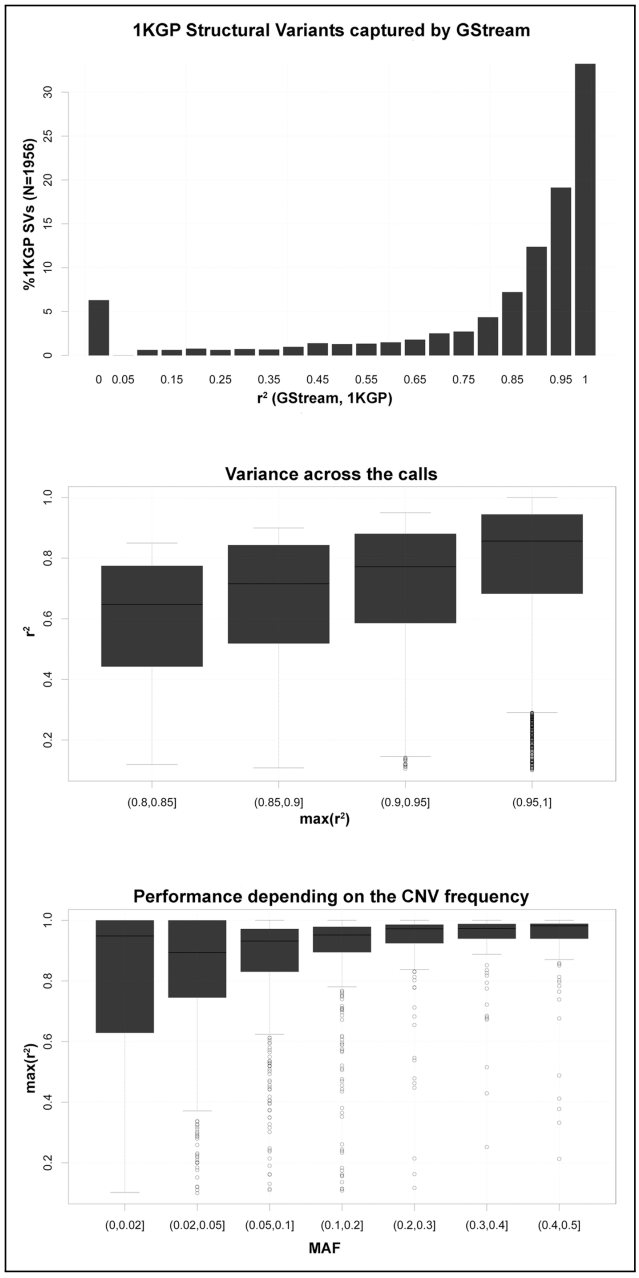


Figure 3.4: *1KGP structural variants captured by GStream.* (A) Percentage of 1KGP structural variants that are captured by GStream within different ranges of r^2 between the 1KGP calls and the GStream calls over the best marker within the respective structural variant loci. (B) Distribution of the r^2 values when more than one marker is found within the structural variant loci. Structural variants are stratified according to the best r^2 obtained by all the markers covering the loci. (C) r^2 distribution stratified by the frequency of the structural variation.

Once demonstrated the power of GStream to capture these variants, we examined the variance of the LD measurements across the markers spanning the same SV loci. This analysis was stratified by the maximum LD measurement of the SV as explained in the previous paragraph. From the results (Figure 3.4B) we can conclude that the calls inferred over probes spanning the same structural variant obtain consistent values with a slight variance due to the quality differences across markers.

Finally, we also observed that the calling performance slightly decreases with the frequency of the analyzed SVs due to an increment of the r^2 interquartile ranges (Figure 3.4C). Nevertheless, lower quartiles exceeded r^2=0.7 within almost all the frequency ranges tested.

We conclude this section by stressing the power of GStream to detect structural variation identified with more advanced technologies (i.e. NGS), obtaining CNV calls with an r^2>0.8 over 75.7% of the 1KGP variants and calls with an r^2>0.9 over 62.3% of the variants.

Genome-wide CNV association study

In order to evaluate the power to detect CNV associations we have performed a pairwise association study between three HapMap populations using golden standard data from three reference studies[24;142;143]. The association statistics obtained by the golden standard calls were compared with those obtained by GStream, CNstream1, PennCNV and QuantiSNP across two microarray platforms, HumanOmni1-Quad and Human1M-Duo. These two platforms were chosen since they represent the first and the second generation of the Infinium HD genotyping microarrays, which mainly differ by the inclusion of additional probes to obtain a better coverage of CNV loci. These differences (Supplementary Figure A.5) are clearly visible and the coverage analysis over the CNV regions defined by the three published studies showed how the marker density is doubled within these CNV regions between the first and the second platform generations (i.e. from 10 to 20 markers/region).

The main metric used to test CNV methods was the $-\log_{10} P-Value$ ratio between the association values obtained by the calls of each method and those obtained from the golden standard calls. Under this metric, ratios near 1 represent a good performance for the method since the obtained association *P*-Value is very similar to the golden standard. A performance summary statistic was computed as the percentage of $-\log_{10} P-Value$ ratios giving values between 0.9 and 1.1 over all the associated CNV loci.

Firstly, the obtained results showed a high performance decrease in the detection of CNV associations when comparing HumanOmni1-Quad results with Human1M-Duo (Table 3.4). This loss was common to all the methods tested but PennCNV and QuantiSNP showed a higher percentage decrease (∼70%) than GStream and CNstream (∼40%). When comparing GStream to the other state-of-the-art methods, results showed a major performance gain

on all the golden standard sets and on both microarray platforms. Results for each platform are described below.

Algorithm	HumanOmni1-Quad				Human1M-Duo				PD[1]
	MCarroll	Campbell	Conrad	Mean	MCarroll	Campbell	Conrad	Mean	
GStream	56.40	46.80	63.87	55.69	45.09	27.65	24.38	32.37	23.3
CNstream	31.11	30.67	30.06	30.61	29.13	17.57	13.89	20.20	10.4
PennCNV	34.66	29.94	29.57	31.39	13.70	9.36	3.73	8.93	22.5
QuantiSNP	33.80	31.50	30.86	32.05	16.50	10.08	4.37	10.32	21.7

[1] Platform difference (PD) refers to the average percentage differences between both platforms.

Table 3.4: Power to detect CNP associations. Percentage of $-\log_{10} P$ ratios higher than 0.9 and lower than 1.1 over the CNV population-associated regions computed for each study.

HumanOmni1-Quad results show that GStream is able to precisely capture an average of 23.6% more associated loci than the other methods (Table 3.4) and that the number of false negatives (Supplementary Figure A.6) decreased considerably. Examining the association ratio distributions (Figure 3.5A), we also observed how GStream outperforms CNstream and, to a greater extent, PennCNV and QuantiSNP: while GStream ratio distribution resembles a unimodal distribution with a high kurtosis and centred around 0.95 (i.e. precisely captured loci), the rest of the distributions from CNstream, PennCNV and QuantiSNP showed a lower kurtosis with more ratio values distributed between 0 and 0.75 (i.e. loci not or poorly captured). To conclude, HumanOmi1-Quad comparison, we examined the association ratio distributions stratified by the *P*-Value obtained by the golden standard calls (Supplementary Figure A.7). GStream obtains very good results within all the *P*-Value ranges, with medians around 1 and decreasing interquartile ranges (from 0.2 to 0.02) as the *P*-Value ranges decreased (i.e. greater evidence of association). Interquartile ranges obtained by the other methods were at least twice the obtained by GStream and increased as the *P*-Value ranges decreased, losing performance when comparing higher associated loci. Poorer ratio medians were also obtained when using the other methods.

When comparing Human1M-Duo results, the performance differences between methods are similar to the previous comparison but with an absolute decrease in performance for all the methods due to previously referred differences on the platform design. Despite this global performance loss, GStream was able to precisely capture three times as many associations as PennCNV and QuantiSNP (Figure 3.5B). The number of associations not detected was also increased by a factor of 3 with respect to HumanOmni1-Quad results (Supplementary Figure A.6). When analyzing $-\log_{10} P-Value$ ratio distributions by their respective golden standard association *P*-Values, ratio medians were only maintained near one when using GStream (Supplementary Figure A.8). Instead CNstream, PennCNV and QuantiSNP showed a significant loss with ratio medians below 0.5.

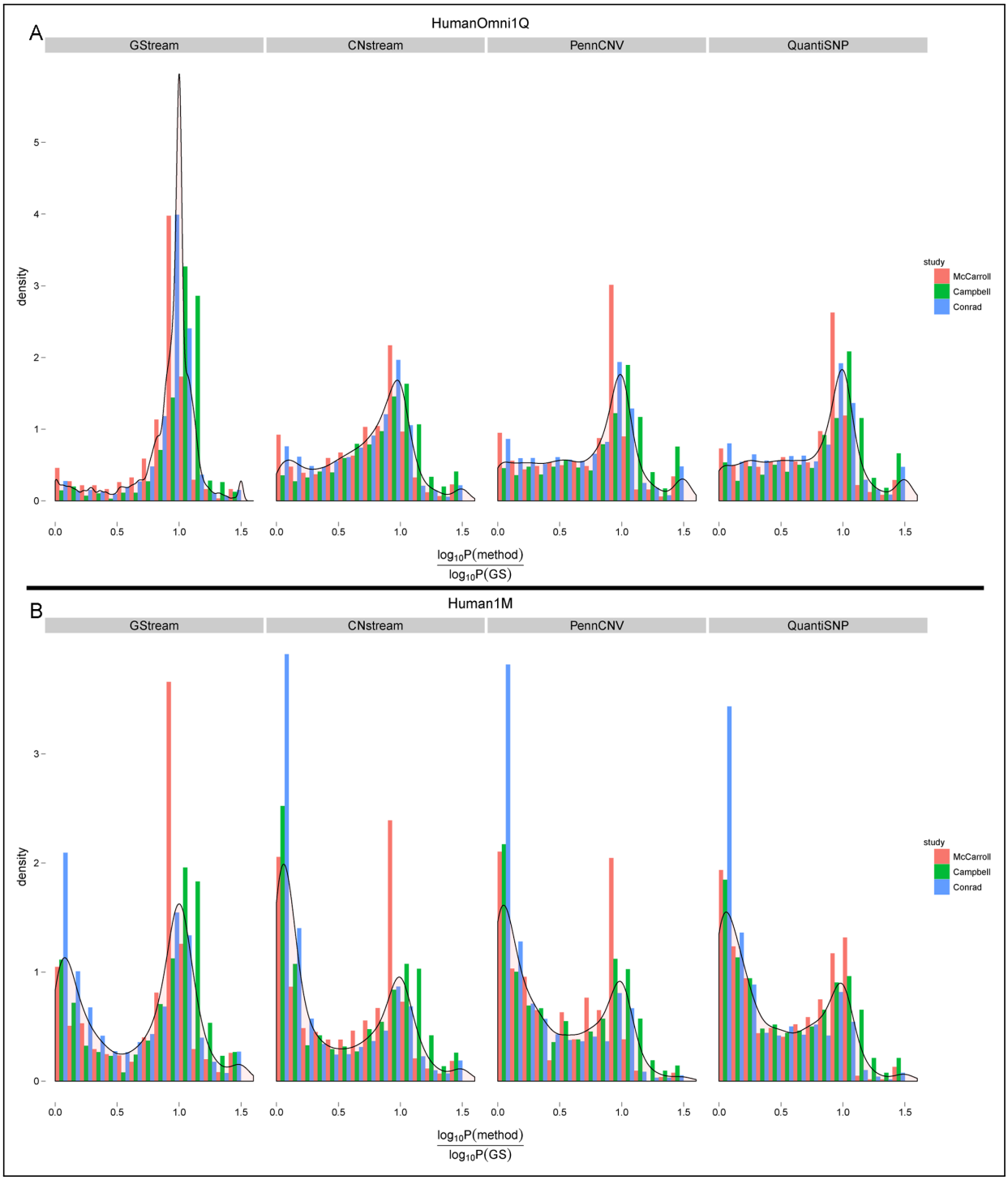


Figure 3.5: *Evaluation of the power to capture genome-wide CNP association.* Plots comparing Chi-square test *P*-Values obtained with the golden standard calls (i.e. McCarroll, Campbell and Conrad datasets) with those obtained with the four tested methods using HumanOmni1-Quad (A) and Human1M-Duo (B) platforms. Comparison is performed by observing the distribution of the *P*-Value association ratios (i.e. tested method versus golden standard). A high performance difference was obtained between the two platforms tested (i.e. due to their high difference in coverage density) and between GStream and the rest of algorithms tested.

3.3.3 Copy number variation and disease susceptibility

Here we describe CNV associations that have been found by mining two available catalogues of disease genomic associations in order to demonstrate the power of GStream to identify new and functionally relevant CNV disease associations.

Catalog of published genome-wide association studies

A set of 333 SNP-CNV pairs have been identified when searching for CNV markers in high LD with trait-associated SNPs reported in the GWAS catalog (Supplementary Table A.1). From this set of paired associations, 94 spanned the HLA region, reflecting the known genomic complexity of this region. On the other hand, previously reported disease-associated CNVs were detected using this approach (Table 3.5 and Supplementary Figure A.9) like, for example, well-known deletions spanning *IRGM*, *LCE3* and *ARMS2* loci, which have been respectively associated to Crohn's disease[155], psoriasis[154] and age-related macular degeneration[158]. A 45 kb deletion near *NEGR1* gene and a 50 kb deletion upstream of *GPRC5B* gene previously associated to obesity and body mass index[159] were also identified.

A thorough study of the CNVs in high LD with trait-associated SNPs revealed several interesting loci. Some of these loci are described below (Table 3.5 and Supplementary Figure A.10):

- A synonymous exonic SNP rs2240335 (*PADI4* gene) has been associated to Rheumatoid Arthritis (RA) in a previous GWAS (P-Value$=2 \cdot 10^{-8}$) using a Japanese cohort (1247 RA cases and 1486 controls)[160]. GStream found 12 CNV markers 2kb away and spanning *PADI4* intron (length = 800bp) in high LD with rs2240335 both on CEU (r^2=0.82) and CHB+JPT (r^2=0.95) HapMap populations.

- An intergenic SNP rs2867125 (50kb downstream *TMEM18* gene) has been associated to body mass index in previous GWAS (P-Value$=3 \cdot 10^{-49}$)[159]. Two CNV loci near this SNP have been identified by GStream in high LD (r^2=1 both on CEU and CHB+JPT populations) with this SNP.

- Levels of glycated hemoglobin have been associated to the *ABCB11* intronic SNP rs552976 (P-Value$=8 \cdot 10^{-18}$)[161]. A deletion variant 3kb upstream of this gene showed a high correlation (r^2=0.94 CEU) with the associated SNP genotypes.

- Intronic SNP rs6815464 on *MAEA* gene has been associated (P-Value$=2 \cdot 10^{-20}$) with type 2 diabetes[162]. This SNP has been found to be highly correlated with another intronic deletion spanning ∼2kb.

- 5'UTR SNP rs6904029 (*HCG9* gene) has been previously associated with Vitiligo (*P*-Value=$1 \cdot 10^{-21}$)[163]. Here we report a CNV locus spanning the 5' region of the *HCG9* gene found in high LD with this SNP.

- Intronic SNP rs3077 (*HLA-DPA1* gene) has been recently associated to chronic Hepatitis B (*P*-Value=$2 \cdot 10^{-61}$) on a Japanese cohort[164]. A deletion potentially spanning three *HLA-DPA1* exons has been identified to be highly correlated with this SNP both on the CHB+JPT and the YRI populations.

- A deletion highly correlated with intronic SNP rs9296736 (*MLIP* gene) has been identified spanning 5kb *MLIP* intron. This SNP has been associated with liver enzyme levels (*P*-Value=$3 \cdot 10^{-9}$)[165].

- Intronic SNP rs2075671 (*ZAN* gene) has been associated to red blood cell count (*P*-Value=$1 \cdot 10^{-9}$) on a GWAS exploring erythrocyte phenotypes[166]. A deletion locus spanning multiple exons of the same gene has been found with a high correlation with the associated SNP genotypes both on the CEU (r^2=0.87) and the CHB+JPT (r^2=0.91) populations.

- Finally, the 3'UTR SNP rs7247513 (*ZNF490* gene), modestly associated with bipolar disorder (*P*-Value=$2 \cdot 10^{-6}$)[167], was also found in high LD with a 2kb deletion covering *ZNF490* intron.

A complete list of all the 333 associations can be consulted on Supplementary Table A.1 and, as previously mentioned, will be regularly updated in our website.

CNV overlap with disease-related genes

In this second approach we examined the CNV variants called by GStream spanning genes known to be involved in disease. A set of 149 CNV consistent loci spanning OMIM genes were obtained (Supplementary Table A.2) with a mean length of 6 kb and a mean number of 7 probes per CNV loci.

In this analysis, well-known associated deletions were found. For example, a common CFH haplotype with deletion of *CFHR1* and *CFHR3* genes associated with lower risk of age-related macular degeneration[168] was identified using the GStream CNV calls. A CNV spanning *CCL3L-CCL4L* genes has been extensively associated with various human immunodeficiency virus-related outcomes [51] and was also identified in this analysis. *SMN*, *GHR* and *PKHD1* gene deletions respectively associated to spinal muscular atrophy[169], responsiveness to growth hormone[170] and polycystic kidney disease[171] were also detected using GStream

(Supplementary Figure A.11). Besides these deletions that have already been associated to disease risk, GStream has also allowed us to identify new exon spanning deletions within disease-associated genes (Supplementary Table A.2). Some examples of these findings are the deletions covering *SLC2A9*, *DAZL* and *MBL2* gene exons.

In almost all these identified CNV loci, GStream calls across the probes within the loci showed a high concordance demonstrating a high performance for a high variety of CNV cluster intensity patterns (Supplementary Figure A.12).

CHR	SNP	SNPbp	CNVbp	N[1]	CEU[2]	YRI[2]	CHBJPT[2]	*P*-Value[3]	Genes	GWAS trait	PMID	Rep.[4]
1	rs2240335	17674537	17677196	12	0.82	0.65	0.95	$2 \cdot 10^{-8}$	PADI4	Rheumatoid arthritis	21505073	No
2	rs2867125	622827	623693	8	1	1	0.94	$3 \cdot 10^{-49}$	TMEM18	Body mass index	20935630	No
2	rs552976	169791438	169776139	5	0.94	0.26	0.32	$8 \cdot 10^{-18}$	ABCB11	Glycated hemoglobin	20858683	No
4	rs6815464	1309901	1290281	3	1	0.38	0.91	$2 \cdot 10^{-20}$	MAEA	Type 2 diabetes	22158537	No
6	rs6904029	29943067	29942384	5	1	1	0.89	$1 \cdot 10^{-21}$	HCG9	Vitiligo	20410501	No
6	rs3077	33033022	33030885	3	0.61	0.94	1	$2 \cdot 10^{-61}$	HLA-DPA1	Hepatitis B	21750111	No
6	rs9296736	53924697	53930407	11	1	1	0.96	$3 \cdot 10^{-9}$	MLIP	Liver enzyme levels	22001757	No
7	rs2075671	100345106	100329189	12	0.87	0	0.91	$1 \cdot 10^{-9}$	ZAN	Red blood cell count	19862010	No
19	rs7247513	12691185	12694963	13	1	1	0.81	$2 \cdot 10^{-6}$	ZNF490	Bipolar disorder	21254220	No
1	rs2568958	72765116	72769429	14	1	1	1	$1 \cdot 10^{-11}$	NEGR1	Body mass index	19079260	Yes
1	rs2568958	72765116	72769429	14	1	1	1	$2 \cdot 10^{-8}$	NEGR1	Weight	19079260	Yes
1	rs2815752	72812440	72769429	14	1	1	1	$2 \cdot 10^{-22}$	NEGR1	Body mass index	20935630	Yes
1	rs4085613	152550018	152557073	46	1	1	0.97	$7 \cdot 10^{-30}$	LCE3-C/D/E	Psoriasis	19169255	Yes
1	rs4112788	152551276	152557073	46	1	0.46	0.97	$3 \cdot 10^{-10}$	LCE3E;LCE3D;LCE3C	Psoriasis	20953190	Yes
5	rs13361189	150223387	150178347	27	1	1	1	$2 \cdot 10^{-10}$	IRGM	Crohns disease	17554261	Yes
5	rs1000113	150240076	150181492	13	1	0.39	0.91	$3 \cdot 10^{-7}$	IRGM	Crohns disease	17554300	Yes
5	rs11747270	150258867	150203780	12	1	0.71	1	$3 \cdot 10^{-16}$	IRGM	Crohns disease	18587394	Yes
5	rs7714584	150270420	150212972	8	1	0.69	1	$8 \cdot 10^{-19}$	IRGM	Crohns disease	21102463	Yes
10	rs10490924	124214448	124217287	10	1	0.94	0.9	$0 \cdot 10^{-0}$	ARMS2	Macular degeneration	21665990	Yes
10	rs3793917	124219275	124216893	10	1	1	0.95	$4 \cdot 10^{-60}$	ARMS2	Macular degeneration	20385819	Yes
10	rs11200638	124220544	124216893	10	1	1	0.95	$8 \cdot 10^{-12}$	ARMS2	Macular degeneration	17053108	Yes
16	rs12444979	19933600	19949684	8	1	1	0	$3 \cdot 10^{-21}$	GPRC5B	Body mass index	20935630	Yes

[1] Number of CNV microarray markers correlated with the SNP genotypes.
[2] Linkage disequilibrium measures (R^2) between the SNP and the CNV in different populations.
[3] Reported GWAS *P*-value.
[4] Indicates if the CNV association has been previously reported.

Table 3.5: CNV loci highly correlated with trait-associated SNPs. This table shows significant SNP-CNV pairs found in high LD.

3.4 Discussion

In this study we present GStream, an integrated tool for SNP and CNV genotyping addressed to Illumina microarray data. This new tool has been carefully designed to obtain a high performance in genotyping accuracy when analyzing GWAS data from Illumina BeadChip arrays. The performance of GStream has been assessed using reference data, extracted from the latest releases of the 1KGP and the HapMap projects, as well as from reference studies on CNV characterization. First, we show that GStream has superior SNP and CNV genotyping performance than current state-of-the-art methods. Second, we demonstrate its power to detect new structural variation recently identified with Next-Generation Sequencing technology. Finally, we also demonstrate the utility of GStream in the identification of CNVs within trait risk loci as well as known disease-associated genes. The newly identified CNV associations could help to advance in the understanding of the genetic basis of several human traits.

In a current scenario where genotyping microarrays are decreasing in cost and widening their spectrum of analyzed SNPs to more rare variations[172], the need of developing methods which increase SNP genotyping accuracy is even more fundamental. To this end, GStream provides a way of facilitating this success by obtaining the best performance results compared to the available state-of-the-art methods (i.e. GenCall[72] and GenoSNP[73]). This increased performance can be particularly meaningful in the case of identifying rare disease-associated SNPs, traditionally more exposed to genotyping errors and to the subsequent statistical bias[173;174]. On the other hand, the accuracy of current SNP imputation methods[175], which expand the number of analyzed SNPs and also help to integrate the results obtained with different microarray platforms (GWAS meta-analysis), also depends on the quality of the originally genotyped SNPs. Therefore, prioritizing accurate SNP genotyping methods is a key success factor in order to obtain reliable imputation results[176].

Besides the importance of SNPs as a source of genetic variation, CNVs have also emerged as important variations for common trait risk[177] as evidenced by recent GWASs[178–182]. In the present study we tested our algorithm power to detect CNV loci that have been recently identified with the Next-Generation Sequencing technology. This NGS CNV data provided by the 1KGP includes not only previously known CNVs (i.e. detected with CGH arrays), but also new CNV loci that have not been previously detected. Since part of these loci are covered by microarray probes, their detection with microarray-based technologies is therefore possible. On the other hand, previous state-of-the-art methods for copy number genotyping[76;77] present a lack of performance when CNVs span a few number of probes or when intensity distributions corresponding to the different copy number states partially

overlap. The multi-component intensity distribution models implemented in GStream will allow researchers to deeply scan the genome for additional CNVs, widening their range to shorter, population-specific and/or previously uncharacterized CNVs.

In this study we also present a two-level comparison of the power of GStream to detect CNV associations in a population-based study. First, we have performed a comparison between the different algorithms tested and, second, we have performed a comparison between the two genotyping microarray generations represented by the Human1M-Duo and HumanOmni1-Quad platforms, this last one including a specific set of markers covering known CNV loci[172]. On the one hand, we confirmed the improvement introduced by the new generation microarrays as a consequence not only of their major density coverage within predefined CNV regions, but also of their improved signal quality. The number of correctly genotyped CNV regions (i.e. characterized in previous reference studies) increased in ∼20% when using HumanOmni1-Quad rather than Human1M-Duo, regardless of the CNV genotyping method being used. On the other hand, when comparing the results obtained by each algorithm tested, GStream showed a higher performance within all the scenarios. Its power increase for detecting and correctly genotyping CNVs (i.e. defined by three different reference studies based either on CGH or custom genotyping arrays) ranges from 50% to 100% compared with the best scoring of the other state-of-the-art methods. Therefore, we present GStream as an integrated SNP-CNV genotyping tool that shows a remarkable leap in performance with respect to previous methods.

One of the most important tasks when analyzing GWAS results is to link the associated variant to a functional effect that can explain the disease risk association. Identifying this link is not always easy since the identified variation can act as a proxy to the underlying causal mutation and may not be covered by the microarray platform. Actually, microarray probe design is based on the study of the linkage disequilibrium patterns and the resulting haplotypes that are inherited in blocks[183]. In this regard, we have identified several CNVs in high LD with SNPs that have been previously associated to disease susceptibility. A clear example of these linked CNVs are the *IRGM1*[155] and the *LCE3B*/*LCE3C*[154] deletions which have been associated to Crohn's Disease and Psoriasis, respectively. Furthermore, these two deletions have been demonstrated to affect the expression of the deleted genes. In addition to these previously known associations, we have identified additional CNVs previously not associated with the disease that could also have functional impact. For example, several CNVs spanning hundreds of bases of gene introns have been found to highly correlate with disease-associated SNPs. These CNVs could provide a functional link to the associated risk modifying, for example, RNA splicing[24;184].

Furthermore, as CNV are known to span from hundreds of bases to multiple kilobases, it is interesting to analyze not only those that correlate previously associated SNPs, but also

those that overlap coding sequences of genes that have been previously associated to human disease (i.e. OMIM genes). The results from this analysis include several known CNV associated loci, as those spanning *CFHR1/CFHR3*[182], *CCL3L/CCL4L*[185], *SMN*[169], *GHR*[170] and *PKHD1*[171] genes. More importantly, new interesting CNV loci also appeared, as those spanning *SLC2A9*, *DAZL* and *MBL2* disease-associated gene exons. The *SLC2A9* gene (OMIM 606142) deletion has been identified by GStream within eight microarray probes spanning two gene exons and two gene introns (chr4:9,929,128-9,966,793). Since mutations within this gene have been previously associated to uric acid concentration[186] and to Hypouricemia[187], the functional effects of this deletion should be further evaluated in relation to these traits. Indeed, GLUT9AN (resulting from alternative splicing of *SLC2A9* gene) is predominantly expressed in the kidney and expression association signals reported by Doring et al.[186] link this gene to the regulation of urate concentrations. The described exon deletion could probably imply a similar effect by modifying the resulting transcribed protein. On the other hand, *DAZL* (OMIM 601486) deletion was identified over 4 microarray probes spanning 5.8 kb (chr3:16,638,525-16,644,130). This deletion could affect multiple gene exons resulting in a drastic functional modification. Previous studies[188] have linked variants within this gene with susceptibility to spermatogenic failure and therefore, this deletion should be evaluated in the context of this human trait. GStream also found a relevant deletion spanning the last exon of the *MBL2* gene. *MBL2* mutations and the consequent Mannose Binding Lectin deficit have been previously associated with cystic fibrosis[189] and recovery from infections[190]. This deletion could drastically modify *MBL2* gene expression and subsequently involve a Mannose Binding Lectin deficit whose association has also been demonstrated.

In a time of rapidly evolving technologies and where Next-Generation Sequencing is becoming available for the study of common diseases, microarray-based technologies are still a commonly used strategy to identify the genetic basis of human traits. First, they allow the analysis of large sample collections at an affordable cost and, second, they have anincreasing global genome coverage, expanding their analysis scope to rarer variants. Therefore, accurate genotyping methods are basic to discover new associated loci that can be then further studied in more detail using Next-Generation Sequencing. The tool that we present in this study, GStream, provides an unprecedented accuracy when analyzing GWAS data from previous and recent Illumina microarray platforms. Furthermore, our software tool implementation allows large-scale GWAS projects to be analyzed in a very short time, providing both SNP and CNV in a single analysis. With these results, we encourage researchers conducting GWAS on these genotyping platforms to use GStream in order to leverage the power of their SNP and CNV loci association analyses.

4 | FOCUS: A Robust Workflow for One-Dimensional NMR Spectral Analysis

Note: This chapter is an exact copy of the paper:
A. Alonso, M.A. Rodríguez, M. Vinaixa, R. Tortosa, X. Correig, A. Julià and S. Marsal. **Focus: A Robust Workflow for One-Dimensional NMR Spectral Analysis**.
Analytical Chemistry, 2014, 86(2), pp 1160-1169. DOI: 10.1021/ac403110u.

Abstract

One-dimensional ^{1}H-NMR represents one of the most commonly used analytical techniques in metabolomic studies. The increase in the number of samples analyzed as well as the technical improvements involving instrumentation and spectral acquisition demand increasingly accurate and efficient high-throughput data processing workflows. We present FOCUS, an integrated and innovative methodology that provides a complete data analysis workflow for one-dimensional NMR-based metabolomics. This tool will allow users to easily obtain a NMR peak feature matrix ready for chemometric analysis as well as metabolite identification scores for each peak that greatly simplify the biological interpretation of the results. The algorithm development has been focused on solving the critical difficulties that appear at each data processing step and that can dramatically affect the quality of the results. As well as method integration, simplicity has been one of the main objectives in FOCUS development, requiring very little user input to perform accurate peak alignment, peak picking and metabolite identification. The new spectral alignment algorithm, RUNAS, allows peak alignment with no need of a reference spectrum and, therefore, it reduces the bias introduced by other alignment approaches. Spectral alignment has been tested against previous methodologies obtaining substantial improvements in the case of moderate or highly unaligned spectra. Metabolite identification has also been significantly improved, using the positional and correlation peak patterns in contrast to a reference metabolite panel. Furthermore, the complete workflow has been tested using NMR datasets from 60 human urine

samples and 120 aqueous liver extracts, reaching a successful identification of 42 metabolites from the two datasets. The open-source software implementation of this methodology is available at *http://www.urr.cat/FOCUS*.

4.1 Introduction

In the last decade metabolomics has experienced an exponential growth thanks to the development and refinement of the analytical techniques used to obtain data from the metabolome[191]. At the same time, advanced bioinformatic methods have emerged in order to deal with the complexity and the high dimensionality of the data generated by these techniques. The recent advances in metabolomics, together with other omic approaches like genomics, transcriptomics and proteomics, are increasingly leading to an integrated knowledge of systems biology and to the consequent understanding of the biologic regulatory processes that underlie metabolism and disease[192–195].

Nuclear Magnetic Resonance (NMR) spectroscopy, together with chromatography-coupled mass spectrometry (MS), are the analytical techniques that have contributed to the growth of the metabolomic science. From the different NMR techniques, the one-dimensional (1D) proton spectrum (^{1}H-NMR) has been the most commonly used technique in NMR-based metabolomics studies[193]. ^{1}H-NMR is characterized by a simple and fast acquisition process[196–198], providing high repeatability spectra that cover a wide range of metabolites[108;197]. However, there are several technical challenges that still need to be solved in order to make the most of this powerful analytical method. These technical challenges affect several of the data processing steps that are required prior to the use of any statistical analysis method. There is actually an intense research activity on solving these technical difficulties, especially on those that more critically affect the quality of the final dataset[122;134;135;196;199–203]. However, while several useful methodologies have been developed for each of the processing steps, there is still a lack of a single tool that efficiently integrates the complete NMR data processing workflow.

One first challenge on the data processing workflow is how to deal with the significant biases in peak positions introduced by the sample chemical environment. Chemical variables like ionic strength, pH and protein content differences between samples can produce changes in peak positions of certain metabolites[108;115]. This means that, even in well-addressed studies with uniform sample processing and randomized designs[204], spectral data analysis will require the application of complex alignment tools in order to minimize the impact of this confounding factors[205]. Multiple alignment methods[117;122–124;206] have been proposed to correct this variability in chemical shifts of NMR spectra and to guarantee peak correspondence across all the analyzed spectra. However, most current methods are still based

on a reference spectrum against which each sample spectrum is aligned. This can be an important source of bias since this reference is not always representative of all the sample spectral diversity[200]. Importantly, methods based on the correlation function[122;123] to compute the alignment correction shifts can introduce other types of errors mainly derived from the presence of very large peaks that account for almost all the weight of the correlation function (despite the presence of lower peaks, where alignment is neglected).

In addition to peak position biases, another major challenge for NMR metabolomics is automatic metabolite identification. MetaboHunter[134] represents an important advance and one of the most common approaches for metabolite identification. In this method, metabolite compounds are matched against a reference spectrum peak list according to the peak positions. However, this approach can lead to a high number of false positives since it does not use intensity and correlation measurements to refine metabolite matching. Another commonly used approach based on the valid cluster concept[135;136] is an improvement with respect to MetaboHunter since, it also uses peak intensities and inter-sample intensity correlation thresholds to group peaks. However, in this method the scores used to match reference spectra to the dataset peaks can lead to suboptimal binary identification when the lowest reference intensity peaks are not found in the sample spectra.

Here, we present FOCUS, a complete workflow for processing large spectral NMR datasets which covers spectral alignment, peak detection, peak quantification and metabolite identification. FOCUS output provides an accurate matrix of sample peak features ready for chemometric analysis, as well as a metabolite identification scoring system that greatly facilitates peak data interpretation. Furthermore, FOCUS generates html-based result reports containing exhaustive information of each analyzed peak such as quality assessment, intensity correlation patterns and metabolite assignations. This integrated tool is suited for the current large sample metabolomic studies and it incorporates several methodological advances both on peak alignment and metabolite identification. FOCUS alignment is based in RUNA (Recursive UNreferenced Alignment) algorithm, which efficiently aligns NMR peaks while avoiding the use of a reference spectrum. With regards to metabolite identification, FOCUS provides a peak-based procedure that significantly speeds up this common and time-consuming task. For each peak, FOCUS generates a list of identification scores for each processed NMR peak, giving the researcher a powerful tool to identify the underlying metabolite.

We demonstrate the improvements of FOCUS in spectral alignment and metabolite identification with respect to other existing methods using both simulated and real (i.e. human urine and liver extract) NMR datasets. With FOCUS, researchers performing moderate to large scale NMR studies will have a complete and powerful tool to make the most of their metabolomic data.

4.2 Theory

FOCUS processes raw 1D-NMR spectral datasets in an integrated and unsupervised manner. As depicted in Figure 4.1, FOCUS analysis pipeline consists of four main steps namely i) Spectral segmentation, ii) Spectral Alignment, iii) Peak detection, and iv) Metabolite Identification. Detailed information on each processing step can be found in the Supplementary Text B.3.

4.2.1 Spectral segmentation

Raw 1D-NMR spectra are automatically divided in equally sized overlapping segments (i.e. 50% overlap) wider enough to span one or multiple peaks (see Supplementary Figure B.1 for further details). Segment overlapping guarantees that peaks falling into one segment end will be correctly analyzed on the consecutive segment. Furthermore, this approach also guarantees that each peak will be analyzed twice within different neighborhoods keeping the best result for further analysis. Subsequent data processing steps concerning alignment and peak picking are performed at the segment-level.

4.2.2 Spectral Alignment

FOCUS alignment method, RUNAS (Recursive Unreferenced Alignment of Spectra), performs spectral 1H-NMR alignment using the cross-correlation function between spectra as the alignment maximization function. In contrast to most of the currently available methodologies, RUNAS does not rely on the definition of a reference spectrum. Furthermore, the alignment procedure is based on a spectral transformation (i.e. Intensity Weight Slope Transform) that enhances peak shapes and reduces the alignment bias produced by the presence of peaks with highly unpaired intensities. RUNAS consists of several processing steps that are outlined below.

First, the Intensity Wight Slope Transform (IWST) is applied to all the input segment spectra. Briefly, this transformation extracts the sign of the spectrum derivative at each point and weights the resulting signal by its corresponding discretized intensity percentile across all the spectrum points. This transformation results in multiple advantages from a spectral alignment standpoint (see Figure 4.2A and Supplementary Figure B.2 for further details). After this first step, RUNAS algorithm proceeds to calculate the optimal distance and the maximal correlation for each pair of samples. The optimal distance is defined as the shift that needs to be applied to one sample spectrum segment in order to maximize its correlation with respect to another sample spectrum. This optimal distance is stored in the optimized distance matrix (ODM). The correlation value between the two spectral segments at

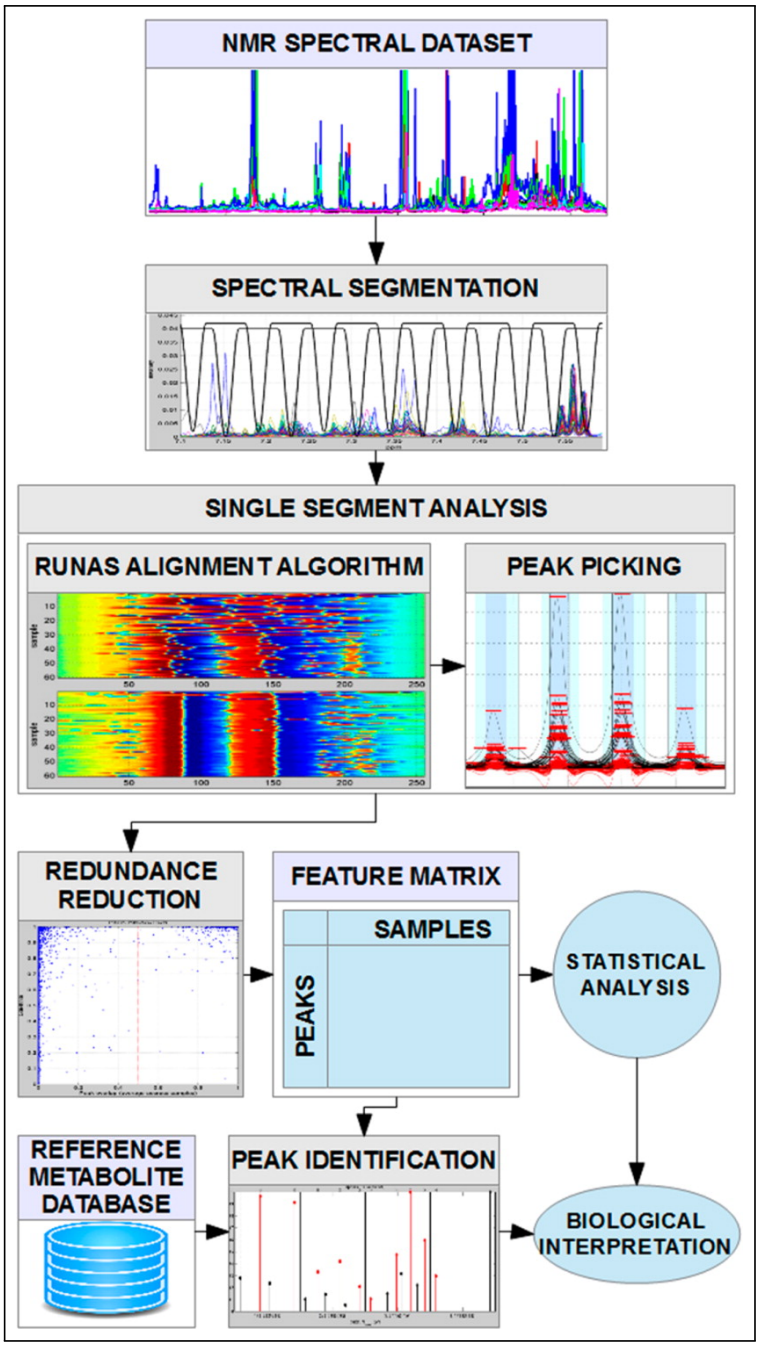


Figure 4.1: *FOCUS workflow schema.* This figure shows the FOCUS workflow analysis for processing one dimensional NMR metabolomics spectra. It starts by computing the informative points of the spectral dataset and then splitting the whole spectra in overlapping segments. For each segment, alignment and peak picking (i.e. peak detection and quantification) are independently applied. Once the segment analysis has finished, redundant peaks are removed and a feature matrix containing peak measurements for each sample is obtained. Statistical analysis can then be performed on this matrix and, after the peak identification procedure, the results from statistical analysis can be interpreted.

this point is called maximal correlation and is stored in the maximized correlation matrix (MCM). All these calculations are based on the Fast Fourier Transform (FFT) [123] and are computed only once. Figure 4.2B shows ODM and MCM matrices for an example segment set

of spectra. Finally, RUNAS performs recursive alignment by iteratively shifting each sample spectrum in order to maximize its spectral correlation with respect to all the other samples. These shifts are calculated as the averaged optimal distances of each spectrum weighted by their corresponding maximal correlation as detailed in equation 4.1:

$$\delta_x^i = \frac{\sum_{y=1}^{N_S} I(\mathrm{MCM}_{xy} \geq C_T)\mathrm{MCM}_{xy}\mathrm{ODM}_{xy}^i}{2\sum_{y=1}^{N_S} I(\mathrm{MCM}_{xy} \geq C_T)\mathrm{MCM}_{xy}}, \forall x \in [1, N_S] \tag{4.1}$$

where x is the spectrum being shifted, i is the algorithm iteration, MCM is the maximized correlation matrix, ODM the optimal distance matrix and $I(c)$ the indicator function whose value is zero unless the comparison c is true, in which case its value is set to 1. C_T refers to a correlation threshold, so spectra that do not reach a minimal correlation with the analysed spectrum are not taken into account in its shift computation. This threshold provides a way to automatically align spectral segments even in those cases where the samples follow more than one peak pattern. At the end of each iteration, the ODM is updated to take into account the applied shifts and the previous step is repeated until convergence. This usually occurs after 10-20 iterations (see Figure 4.2C and Supplementary Figure B.3). Figure 2D shows how the example segment set of spectra is correctly aligned without using a reference spectrum and how groups of different spectra are independently aligned.

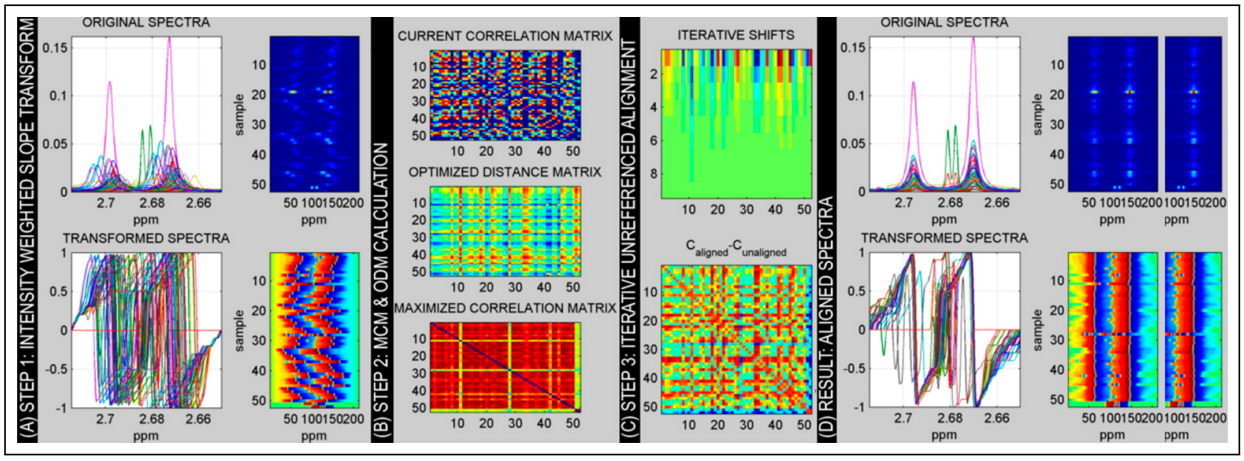


Figure 4.2: *FOCUS spectral alignment.* This figure shows the FOCUS procedure for spectral alignment. (A) Shows the spectral signal transformation (intensity weighted slope transform) and how signals and peaks are equalized across samples by using this transformation. (B) Shows the resulting MDM and ODM matrices computed before recursive alignment. (C) Shows the matrix of applied shifts (rows representing iterations and columns applied shifts on each sample) and the difference between the final and the initial spectral correlation matrices. (D) Shows the aligned spectra after applying FOCUS. It can be observed that each signal is only aligned against those signals achieving a high degree of spectral correlation.

4.2.3 Peak Detection

FOCUS peak detection approach is based on the computation of a consensus peak signal (CPS) which estimates at each spectral point the frequency of samples having a peak region spanning the considered point. For each sample spectrum, the zero-crossings of the filtered spectrum (i.e. second derivative Gaussian filter) are used to delimitate peak regions on each sample. The CPS is then built by computing at each spectral point the number of samples that present a peak region spanning the considered point and scaled with respect to the number of samples. The CPS is then an estimation of the peak frequency across the samples at each spectral point. In this way, this method guarantees that all the samples will have the same contribution to the definition of peak regions. Once the CPS has been computed it is filtered using the same previous filter and the resulting signal zero-crossings delimitate the global peak regions (see Supplementary Figure B.4A for further details). Using CPS to delimitate peak regions also allows to take into account the residual misalignment that may remain after peak alignment, since the residual variability on the single-spectrum peak regions will contribute to broaden the corresponding peak on the CPS (see Supplementary Figure B.4B).

Since input spectra are divided in consecutive overlapping segments, redundant peaks can be generated. In order to remove this redundancy a peak reduction method based on the sample-averaged overlapping within peak pairs extracted from consecutive windows is applied. Peak pairs having large sample-averaged overlapping correspond to redundant peaks and the method keeps only those peaks with larger peak shape correlations.

At this point, FOCUS provides a data matrix of per-sample peak measurements that is ready for subsequent chemometric analysis. The user can select different types of per-sample peak features, namely i) peak areas, ii) peak maximums, and iii) peak increments (difference between the peak maximum and minimum value at the peak region limits). Furthermore, three quality scores are provided for each peak, namely i) sample-averaged peak shape correlation (i.e. this can be used to evaluate the peak shape correspondence between samples), ii) correlation between areas and increments (i.e. this can be useful for detecting background interferences or overlapping peaks which might be deconvoluted) and iii) the median peak intensity percentile with respect to all the detected peaks (i.e. this quality score provides information about the relative concentration level of the metabolite that generates each peak).

4.2.4 Reference-Based Metabolite Identification

FOCUS reference-based metabolite identification is based on the fact that closer NMR peaks in one pure compound spectrum arise from the same proton and they use to show less variation across the frequency axis than peaks generated from different protons. Consequently,

given a library of known metabolites (i.e. reference spectra) FOCUS will start by clustering the reference peaks of each metabolite proton (see Figure 4.3A). In this way, each proton cluster will group together peaks with close positions in the spectrum (i.e. distance between peaks $< d_{cluster}$). After this clustering step, a peak-oriented identification step is performed, where the set of target metabolites for each dataset peak is limited to those having a close reference peak (i.e. distance between dataset peak and reference peak $< t_{cluster}$, see Figure 4.3B). For each dataset peak and a corresponding target metabolite, FOCUS proceeds as follows:

1. Intra-cluster score: This step consists of identifying the metabolite peaks of the same proton. For this objective, FOCUS tries to identify correlated dataset peaks (i.e. intensity correlation at the sample level) at the expected positions where the other peaks from the same proton should be located (see Figure 4.3C for further details). If a peak from the same proton is not identified, a zero-intensity peak is assigned to it. Finally, the correlation between the intensity levels of the correlated dataset peaks and their respective reference peaks is computed in order to obtain the intra-cluster matching score.

2. Inter-cluster score: This step consists of identifying the metabolite peaks of the other reference metabolite protons (see Figure 4.3D). The inter-cluster matching score is computed as the number of protons of the target metabolite where correlated dataset peaks have been found, and scaled by the total number of protons. In this calculation, the weight of each proton depends on the intensity of its reference peaks (i.e. protons having relevant peaks will have more importance).

3. Penalization score: If correlated dataset peaks are found outside the windows defined around the peaks of the metabolite protons, a penalization score is computed as $Sp = 1 - wN$, where N is the number of correlated peaks outside the windows and w a user defined weighting factor that establishes the degree of penalization.

Following this procedure, FOCUS obtains, for each detected peak, a list of candidate metabolite identifications. This candidate list can be sorted by the corresponding identification scores (i.e. average of the intra-cluster, inter-cluster and penalization scores) in order to identify the metabolite that better represents this dataset peak. If a reference metabolite has only one peak, the scoring is performed by averaging the penalization score with a singlet-score, which is proportional to the closeness of the reference and dataset peaks. After this identification step, FOCUS also creates a putative annotated feature matrix, linking each peak with its top-scored metabolite identification.

4.2.5 Results Report Generation

In addition to the per-sample peak feature matrix, FOCUS generates an exhaustive and user-friendly results report which greatly facilitates the exploration of the obtained results. This web-based report is created locally (i.e. no need of accessing an external server) and uses JavaScript plugins to improve user navigation along the results. For each analysis, a summary report is created with the list of detected peaks characterized by their positions, their quality scores and the best metabolite identification. A set of group identifiers are also provided, where each peak group refers to peaks that have been grouped due to their high intensity correlations. Besides this summary report, each peak contains a link to another webpage that contains detailed information such as correlated peaks and all the metabolites that can correspond to the peak sorted by their identification scores. For further details see Supplementary Figure B.5 and the example report available on the methodology webpage (*http://www.urr.cat/FOCUS*).

4.2.6 Software

The FOCUS methodology presented here has been developed using Matlab (i.e. data processing) and Python (i.e. automatic report generation). The Matlab package and a demo example can be downloaded from *http://www.urr.cat/FOCUS*. A Python script for creating a web-based report of the results generated by FOCUS and an example report can also be accessed from this web.

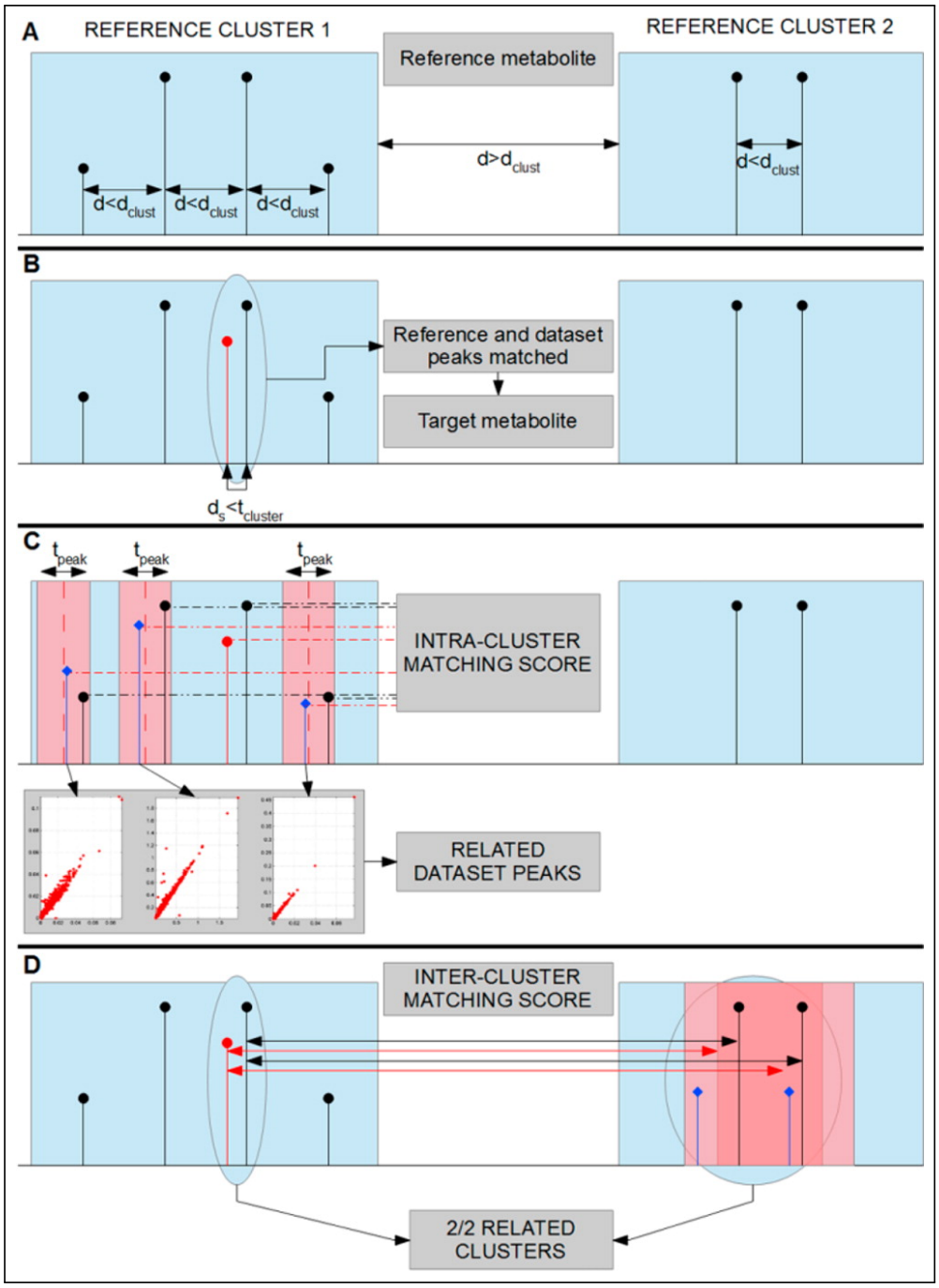


Figure 4.3: *Metabolite identification algorithm.* This figure shows the processing steps for metabolite identification. (A) The peaks of each reference metabolite spectrum are grouped in clusters (i.e. putative protons) if their distance does not exceed the maximum intra-cluster distance $d_{cluster}$. (B) The reference metabolite of subfigure A is a target metabolite for the dataset peak (red line) since one of its peaks (black lines) is close to the dataset peak (red line). (C) Intra-cluster matching is performed by identifying correlated dataset peaks (i.e. blue lines) at the positions where the other reference proton peaks should be found (a tolerance window t_{peak} is defined around each expected position). Scatter plots show the intensities of the correlated peaks (blue peaks) against the intensity of the peak being identified (red peak). Finally, the intra-cluster score is defined as the correlation of the mean intensity levels of the dataset peaks (blue and red peaks) versus the library defined intensities of the target metabolite peaks (black peaks). (D) Inter-cluster matching is defined by the number of protons having at least a dataset related peak. Search windows are defined by the reference cluster distances and expanded by a tolerance factor.

4.3 Experimental section

4.3.1 Liver extract dataset

Liver extract ^{1}H-NMR measurements of 120 samples were conducted as previously described. Briefly, for each hepatic sample, 50 mg of tissue was removed, flash-frozen, and mechanically homogenized in 2 mL of $H_2O/CH_3CN(1/1)$. Each homogenate was centrifuged at 5000g for 15 min at 4°C and the supernatant containing hydrophilic metabolites was subsequently frozen at -80°C. For NMR measurements, the hydrophilic extracts were reconstituted in 600 μL D_2O containing 0.5 mM trisilylpropionic acid (TSP) and transferred to a 5 mm NMR tube. The 1D Nuclear Overhauser Effect Spectroscopy with a spoil gradient (NOESY) was used to record 1D ^{1}H-NMR spectra using a 600.2MHz frequency Avance III-600 Bruker spectrometer (Bruker, Germany) equipped with an inverse TCI 5 mm cryoprobe. A total of 256 transients were collected across 12kHz spectral width at 300K into 64k data points, and exponential line broadening of 0.3 Hz was applied before Fourier transformation. A recycling delay time of 8s was applied between scans to ensure correct quantification. The frequency spectra were phased, baseline corrected and then calibrated (TSP, δ=0.0ppm) using TopSpin software (version 2.1, Bruker, Germany).

4.3.2 Human urine dataset

Control individuals were recruited as part of the Immune-Mediated Inflammatory Disease Consortium (IMIDC)[207] repository. Urine samples were obtained from 60 healthy individuals attending to blood bank donations at university hospitals from different regions in Spain in collaboration with the Spanish National DNA Bank. Urine was collected into 15 ml universal containers with clorhidric acid. All samples were processed 24 hours later, and stored at -80°C until analysis. After thawing, 400μL of each HCl-preserved urine sample were aliquoted and we added 200μL of buffer phosphate (1.5 mM $Na_2HPO_4/NaH2_PO_4$ in D_2O with 0.62 mM of TSP, pH=6.8) to each sample. The final solution was transferred to a 5 mm NMR tubes for subsequent ^{1}H-NMR acquisition. We applied the same parameters described above for spectral acquisition and processing. However, in this case the recycling time was set to 7s.

4.3.3 Metabolite databases

When running FOCUS on the ^{1}H-NMR liver extract and human urine datasets peaks were annotated by using in-house metabolite databases (see Supplementary Table B.1). These databases contain the peak positions and relative intensities of the metabolite reference spectrum measured at pH=7.5.

The database used for metabolite identification in the liver extract dataset contains 31 reference metabolite spectra and is based on a list of previously manually identified metabolites. The database used in the human urine dataset analysis was composed of 47 reference metabolite spectra that are commonly found within this biofluid[141;196] either as endogenous/exogenous metabolites or as potential sample handling contaminants.

4.3.4 Alignment Performance Evaluation

We have tested FOCUS alignment performance and compared it to two of the most commonly used alignment methods: Icoshift[122] and Correlation Optimized Warping (COW)[123]. Icoshift algorithm is based on segment-shifting by maximizing spectral correlation against a reference signal. Like FOCUS, Icoshift uses the FFT in order to speed up calculation as proposed by Wong et al.[123]. The reference spectrum is commonly computed as the average spectrum of all the sample data and the segments in which the whole spectra are divided can be specified by the user or otherwise determined to be regularly spaced. COW is a warping alignment method based on dynamic programming that stretches or compresses segment sections in order to better match the signals to be aligned. This method is also based on a target spectrum against which sample spectra are aligned. The input data for this method consists of the segment length and the maximum warping range in the specified segment length.

In order to evaluate and compare these alignment methods, we first used simulated spectral datasets under a large range of parametric scenarios. These datasets were characterized by the presence of two and three peaks per sample, using the Lorentzian distribution[108] as the basis function for building simulated peaks (see Supplementary Text B.4 and Supplementary Figure B.6 for details) and the different parametric scenarios have been taken into account by jointly modifying the following parameters: (a) distance between peaks, (b) standard deviation of the applied shifts to unalign sample spectra, (c) scale parameter of the Lorentzian distribution, (d) peak missingness, (e) sample size, and (f), intensity ratio between peaks. Evaluation was performed on 973 parametric scenarios for each dataset considering all the possible permutations. The first metric used to evaluate alignment performance was the distance between the originally aligned spectra and the algorithmically aligned spectra correlation matrices. The second metric was the correlation between the known shifts applied to unalign the dataset and the shifts derived from the algorithmic alignment. This metric was only computed for FOCUS and Icoshift, since COW approach does not use alignment shifts.

In order to also test the alignment performance over a real dataset, we have also used a human urine NMR dataset composed of 60 samples where spectral unalignment due to pH

inter-sample variation is clearly visible. The performance was measured using the averaged spectra correlation[208] on different sample sizes (i.e. 10, 30 and 60 samples). A total of 48 informative segments were selected across the entire spectrum to perform this analysis in order to avoid performance evaluation over uninformative segments (see Supplementary Figure B.7).

4.4 Results and discussion

4.4.1 Alignment Performance Evaluation

Spectral alignment on the simulated spectral datasets showed a significant improvement when using FOCUS in comparison to Icoshift and COW (see Figure 4.4 and Table 4.1). Applying FOCUS significantly reduced the correlation matrix distance between the true aligned spectra and the algorithmic aligned spectra when compared with Icoshift (Wilcoxon test; *P*-value=$1e^{-41}$) and COW (Wilcoxon test; *P*-value=$8e^{-38}$). FOCUS improvements can also be extended to the shift correlation performance measures (see Table 4.1), where FOCUS improved the Icoshift alignment results both on the doublet dataset (i.e. 10%) and on the triplet dataset (i.e. 5%) when evaluating the distances between the expected and the obtained correlation matrices. A more detailed analysis of the results (see Supplementary Figures B.8, B.9 and B.10) shows that COW performance was more sensitive to sample shifts than FOCUS and Icoshift, being FOCUS the algorithm with the most stable performance against changes in the degree of spectral unalignment, the distance between peaks, peak width and sample size. The simulation analysis results showed the robustness of FOCUS method against high degrees of unalignment, given that the correlation matrix distances only increased in 2.5% (i.e. from $1.63e^{-3}$ to $1.67e^{-3}$) when doubling the standard deviation of the applied shifts to unalign the spectra. Instead, Icoshift and COW, showed increases of 19.7% (i.e. from $2.71e^{-3}$ to $3.24e^{-3}$) and 76.9% (i.e. from $2.11e^{-3}$ to $3.74e^{-3}$), respectively. Therefore, the RUNAS alignment algorithm implemented in FOCUS shows clearly a better performance than the previous methodologies, making it particularly suitable to those datasets suffering moderate to high unalignment bias either globally or in specific spectral regions.

When evaluating spectral alignment results on the human urine spectral dataset we also found a significant performance improvement when using FOCUS (see Figure 4.5A and Table 4.1). FOCUS, Icoshift and COW respectively showed average spectral correlation improvements of 53.49% (from an initial correlation of 0.46 to 0.70), 39.38% (from 0.46 to 0.64) and 35.64% (from 0.46 to 0.62) with respect to the unaligned spectral dataset. Reducing the number of analyzed samples (i.e. from 60 to 30 or 10 samples) produced

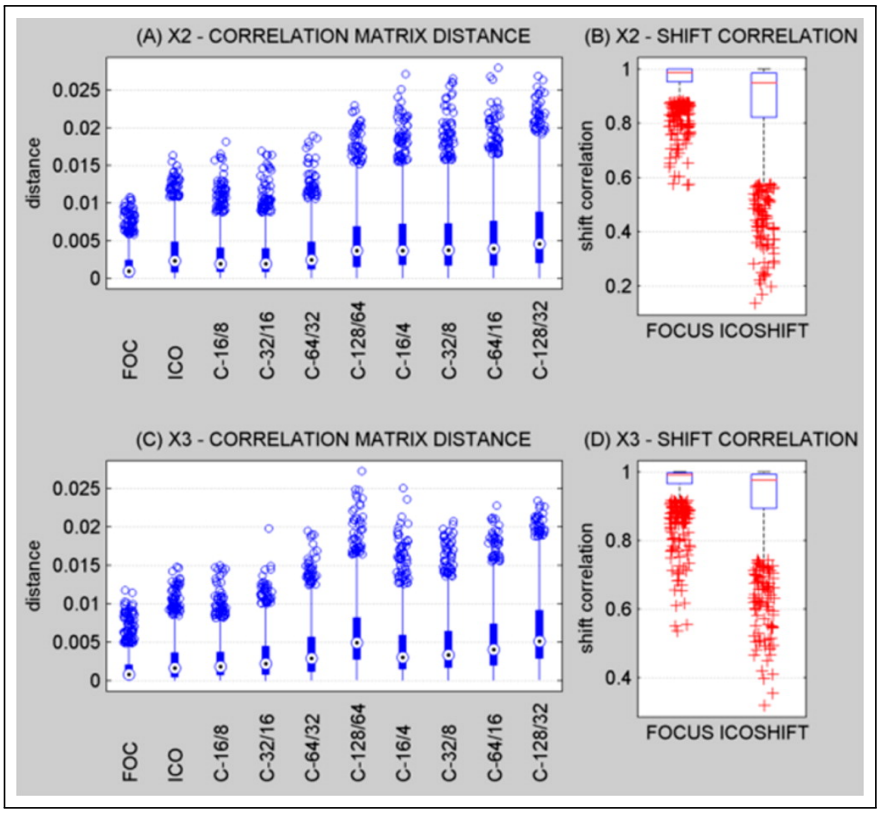


Figure 4.4: *Simulated datasets alignment results.* This figure shows the alignment performance measured over the simulated datasets. (A) and (C) show the distribution of distances between real and algorithmically aligned spectral correlation matrices across the different parametric scenarios over the doublet (X2) and triplet (X3) datasets. In the same way, (B) and (D) show the distribution of correlation coefficients between the true and the algorithmically computed shifts by FOCUS and Icoshift. COW method was evaluated using a combination of different parameters (C-[segment length]/[slack][117]).

only moderate performance reduction in the three methods, with consistently better results with the FOCUS aligned dataset. When regarding per-sample correlation results, FOCUS improvements were similar in all the spectra (see Supplementary Figure B.11). This confirms that FOCUS performance is superior in any sample size and does not depend on the particular sample spectra. The effects of applying FOCUS on unaligned spectral datasets are clearly visible on those segments which peaks are more susceptible to suffer chemical shifts. Figure 4.5B shows a spectral segment corresponding to hippuric acid peaks that are highly affected by unalignment. In these cases, the average spectrum clearly does not represent the true peak distributions. FOCUS superior performance cannot only be derived from its correlation gain but also from the increased height to width ratio of the resulting peaks on the averaged spectrum and from the removal of peak artifacts introduced by the other alignment methodologies (see Figure 4.5B).

4.4.2 Automated Analysis of the Human Urine Dataset

FOCUS processing workflow was applied to a set of 60 human urine spectra. The unsupervised analysis was performed using the default parameter values: windows with a 50%

Algorithm	Simulated dataset				Human Urine dataset
	Distance(C)[1]		C(Shift)[2]		C(Spectra)[3]
	X2	X3	X2	X3	
FOCUS	1.7e-3	1.6e-3	0.96	0.97	0.70
ICOSHIFT	3.4e-3	2.7e-3	0.87	0.92	0.64
COW	3.0e-3	2.8e-3	na	na	0.62

[1] Distance between the true and the algorithmically aligned correlation matrices.
[2] Shift correlation between the true applied shifts and the correction shifts applied by the algorithms.
[3] Averaged spectra correlation.

Table 4.1: Performance alignment results. This table shows the most important performance measures obtained within the simulated and human urine datasets. Two simulated datasets were evaluated respectively presenting two (X2) or three (X3) peaks per sample.

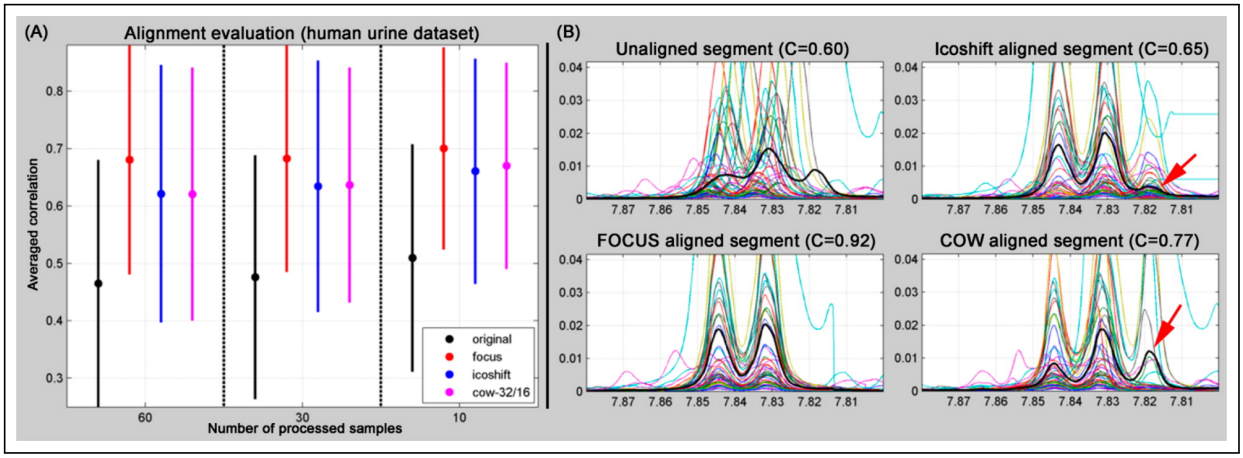


Figure 4.5: *Human urine alignment results.* This figure shows the alignment performance measured over a dataset of 60 human urine NMR spectra. (A) shows the averaged spectra correlation before (i.e. original) and after alignment (i.e. FOCUS, Icoshift and COW). These performance measures were computed on the complete dataset (60 samples) but also reducing the number of analyzed samples (i.e. 30 and 10). (B) shows a spectral segment corresponding to hippuric acid peaks and the alignment results for the three algorithms tested. Black lines represent the averaged spectrum. These figures show how unalignment can introduce spurious peaks that are not solved by Icoshift and COW algorithms (red arrows).

of overlap and a length of 0.077 ppms (256 spectral data points), minimum peak width was of 0.01 ppms and minimum sample frequency to consider a peak was 10%. The only parameter that needed to be adapted was the minimum intensity increment to consider a peak which depends on the intensity scale of the analyzed spectra. This minimal parameter adaptation reflects one of the FOCUS strengths, which is its capacity to easily adapt to the singular characteristics of different NMR datasets with very few user inputs.

After the sliding window analysis, a set of 390 peaks were obtained which was reduced to 240 peaks (i.e. 38% reduction) after applying redundancy reduction. The mean intensity correlation between the redundant peaks was of 0.948 which confirms the accuracy of the

redundancy reduction approach (see Supplementary Figure B.12). At this point, the NMR data processed by FOCUS represents a quality set of per-sample peak measurements that can be directly used to perform statistical analyses.

The correlation analysis identified 188, 168, 142 and 123 peak groups using the default grouping correlation thresholds of 0.95, 0.90, 0.85 and 0.80, respectively. Regardless of the correlation threshold used, 20 peak groups gathered more than two peaks, being clear candidates for the identification step since their peaks were highly correlated between them and uncorrelated with the remaining dataset peaks. These correlation patterns between peaks are used by the FOCUS metabolite identification algorithm to assign to each dataset peak the most plausible metabolite. Metabolite clusters were determined using the default maximum intra-cluster distance of 0.05 ppm and the intensity correlation threshold to consider related dataset peaks was set to 0.80. The tolerance window for matching reference and dataset peaks and for inter-cluster peak matching was set to 0.03 ppm thus allowing for a certain shift of the dataset peaks with respect to the reference metabolite peaks. Finally, the peak tolerance window for intra-cluster matching was set to its default value (0.005 ppm) given that the distance between peaks of the same cluster does not depend on pH variations. This procedure obtained a set of 22 correct metabolite identifications that were manually verified and supported by 43 peak-metabolite associations (see Table 4.2 and Supplementary Table B.2). These identifications demonstrated the good performance of FOCUS identification procedure. For example, peaks associated to citrate, hippurate and trigonelline obtained large identification scores (i.e. respectively 1.00, 1.00 and 0.97) due to the presence of correlated peaks inside each cluster and between the clusters (see hippurate identification example on Figure 4.6). Creatinine is also another example of correct identification based only on the inter-cluster matching since reference clusters are only composed of one peak (see Supplementary Figure B.13). Furthermore, other identifications as lactate also obtained large identification scores (i.e. 0.94) although clusters with lower intensity peaks were not identified, since FOCUS prioritizes the identification of clusters with higher intensity peaks (see Supplementary Figure B.13).

4.4.3 Automated Analysis of the Liver Extracts Dataset

FOCUS processing workflow was also applied to a set of 120 liver extract NMR spectra. Like in the previous analysis, only the minimum intensity increment was changed with respect to the default values. After alignment and peak detection steps, a set of 413 peaks was obtained which was reduced to 228 (i.e. 44.8% reduction) after the redundancy reduction step. The averaged intensity correlation between the redundant peaks was of 0.995 also confirming the accuracy of the approach (see Supplementary Figure B.11).

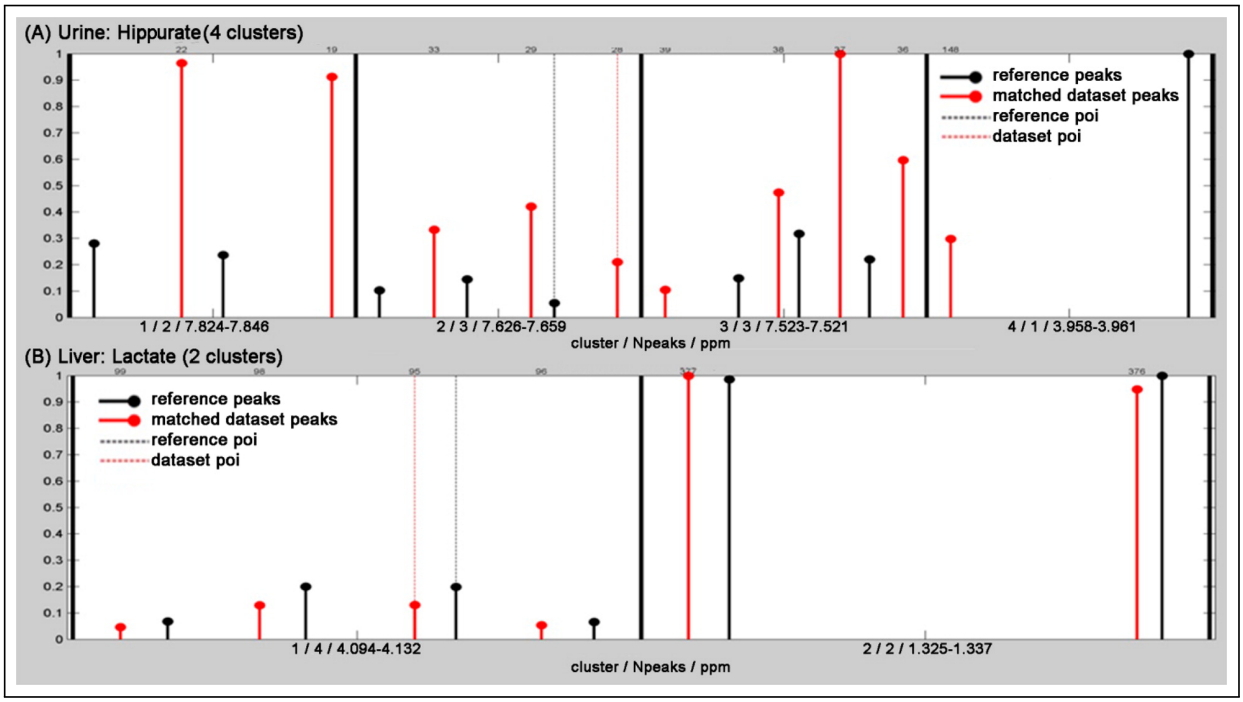


Figure 4.6: *Metabolite identification.* This figure shows two examples of successful identification of metabolites on NMR datasets using FOCUS. (A) Shows the identification results of hippurate on the human urine dataset. Hippurate reference spectrum is characterized by four clusters (clusters are separated by thick black lines and reference peaks are represented with black lines ended with circles proportional to the reference peak intensities). Dataset peaks matched are represented with red lines. (B) Shows the identification result of lactate on the liver extracts dataset.

The correlation analysis provided 112, 81, 53 and 36 peak groups using the default grouping correlation thresholds (0.95, 0.90, 0.85 and 0.80). The number of groups gathering more than two peaks was 15 at all the correlation thresholds, then being good candidates for the identification step. Since this dataset shows a lower peak position variance across samples, the tolerance window for matching reference and dataset peaks and for inter-cluster peak matching could be reduced down to 0.01 (in comparison to the tolerance parameter used in the human urine dataset which was of 0.03). For the same reason, grouping correlation threshold was increased up to 0.90. The metabolite identification procedure succeeded to correctly identify 20 metabolites on this dataset supported by 63 peak-metabolite associations rightly matched (see Table 4.2). Metabolite compounds like lactate, taurine, tyrosine and glucose achieved high identification scores due to high intra-cluster and inter-cluster matching scores (see lactate example on Figure 4.6). Lactate identification is a clear example of how FOCUS is able to discriminate identification of partially overlapping metabolites by detecting non-overlapped multiplets: in this case, both lactate and threonine have an identical doublet on the 1.33 ppm range. Nevertheless, the identification of correlated lactate peaks from another multiplet (4.08 ppm) increased the lactate score in comparison to the threonine score (i.e. threonine related multiplets were not found). Other identifi-

Liver Extracts Dataset			Urine Dataset		
Metabolite	Score	ppm	Metabolite	Score	ppm
Creatine	1	3.041	Citrate	1	2.692
Glucose	1	4.642	Creatinine	1	3.054
Lactate	1	4.118	Hippurate	1	7.846
Taurine	1	3.437	Dimethylamine	0.97	2.72
Acetate	0.99	1.92	Trigonelline	0.97	4.448
Choline	0.97	3.206	Choline	0.96	3.209
Fumarate	0.97	6.521	Creatine	0.96	3.939
Alanine	0.95	1.489	Alanine	0.95	1.496
Leucine	0.95	0.978	Lactate	0.94	1.347
Isoleucine	0.93	1.01	Ethanol	0.93	1.203
Tyrosine	0.92	7.205	Formate	0.93	8.468
Glycerol	0.9	3.647	Methanol	0.93	3.367
Valine	0.9	1.041	Taurine	0.92	3.418
β-hidroxybutyrate	0.88	1.208	Acetone	0.88	2.241
Glutamine	0.86	2.463	Acetylcarnitine	0.86	3.201
Creatinine	0.85	3.057	TMAO	0.85	3.288
Uridine	0.85	7.871	Glycine	0.83	3.577
ADP	0.8	8.233	Methylguanidine	0.78	2.832
Phenylalanine	0.78	7.432	Pyruvate	0.74	2.348
Glutamate	0.73	2.36	Cis-aconitate	0.74	3.107
			Acetate	0.68	1.931
			Phenylalanine	0.66	7.43

Table 4.2: Metabolite identification results. This table shows the successful metabolite identifications on the urine and the liver extracts datasets using FOCUS.

cations, like glutamine, are good examples of the algorithmic improvements achieved by FOCUS intra-cluster matching. This metabolite has a peak cluster between 2.39 and 2.54 ppm that groups 12 peaks. From these 12 peaks only 4 peaks have been found in this dataset with high intensity correlations. This low ratio of cluster identified peaks would result in low identification scores in previous methodologies. Instead, FOCUS obtains a high identification score (i.e. 0.84), since the reference peaks that have not been identified are characterized by low intensities with respect to the identified peaks also overlapping with glutamate peaks (see Supplementary Figure B.14). The same can be observed on the phenylalanine cluster between 7.32 and 7.44 ppm, where only 4 of 8 peaks were identified but corresponded to the highest intensity reference peaks and showed the same intensity pattern. Importantly, the identification algorithm robustness against peak position shifts can be clearly observed in the lactate and the taurine identifications (see Figure 4.6 and Supplementary Figure B.14). Lactate peaks on the liver spectra were found with a uniform shift of 0.003 ppm with respect to the reference. The first taurine peak cluster showed a perfect matching between reference and dataset peaks, while the second cluster showed a

difference of 0.008 ppm.

Although multiple peaks have been unequivocally associated to one metabolite, other peaks may be associated to more than one metabolite with high identification scores. Such is the case of peak 363 at 1.754 ppm (see Supplementary Table B.3), which is located in the overlapping region of lysine and leucine spectra and obtains identification scores of 0.83 and 0.82 for lysine and leucine, respectively. In such cases, the exploration of the correlation patterns of the peaks of interest or the identification of non-overlapping multiplets can help to unequivocally choose the optimal metabolite (i.e. leucine identification is reached by identifying its triplet on 0.948 ppm). FOCUS provides an interactive tool which will allow the users to easily navigate on the generated results for each peak, significantly facilitating this time-consuming task.

4.4.4 Discussion on general algorithmic performance and analytical technique limitations

Our results show a good performance of FOCUS in both synthetic and real datasets, thus making it a suitable solution for untargeted profiling of large-scale metabolomics ^{1}H-NMR spectra studies. FOCUS has demonstrated to accurately handle spectral artifacts and the most common analytical source of variances of this type of profiling studies.

Automated segmentation analysis with overlapping avoids peak detection errors when the sample-spectrum peaks are close to the bounds of the analyzed segment, since they will be close to the middle of the following segment. Although FOCUS has demonstrated a very good performance handling misaligned spectra, in some rare occasions slight misalignment errors have been observed for specific segments containing peaks with uncorrelated shifting patterns. In such cases, FOCUS provides two complementary methods to reduce the potential impact of this limitation. First, the overlapping segmentation analysis guarantees that each peak will be analyzed twice with different neighborhoods, keeping only the best alignment result. Second, the FOCUS peak detection method based on the CPS sets the peak integration range accounting for the residual shift variability that may have not been corrected by the peak alignment method. In those rare cases where these methods are not sufficient to deal with these problems, the users can choose to set their own limits around the peak of interest to avoid the interferences of the neighboring peaks.

FOCUS peak detection method has also shown to have a large dynamic range (i.e. 1:1000) on the analyzed datasets. This is due to the use of a frequency threshold: true signal peak positions are correlated while noise peaks are uncorrelated so they will rarely exceed the frequency threshold. As previously commented, CPS signal has also demonstrated to be

robust against residual shifts by broadening the integration area (see Supplementary Figure B.4B).

The identification module in FOCUS attempts to assign metabolites to NMR peaks according to a database of standard spectral references. FOCUS has been able to properly identify most of the common metabolites present in the studied NMR spectra from biological samples. Furthermore, FOCUS identification report includes several quality control measurements aimed at facilitating the identification of those cases where metabolite identity assignment can be difficult due to the inherent technical limitations of an untargeted analysis (i.e., heavily overlapped signals).

4.5 Conclusions

The results presented here show that FOCUS NMR analysis software addresses the main problems that still affect to the processing of high-throughput NMR metabolomic data, FOCUS provides an integrated workflow that performs all the necessary processing steps required to obtain a set of spectral measurements ready for the usual chemometrics analyses. Importantly, the FOCUS algorithm for spectral alignment (i.e. RUNAS) has demonstrated a highly significant improvement when dealing with moderate to highly unaligned spectral datasets. This behavior has been achieved by applying an innovative approach where no reference spectra are needed and the raw spectra are mathematically transformed to maximize peak alignment and minimize outlier artifacts. Furthermore, the peak detection method also avoids the bias produced by outlier samples and bases its detection on the peak-sample frequency. Additionally, FOCUS also provides a new and efficient method for metabolite identification, facilitating this time-consuming task.

In order to demonstrate the usefulness of FOCUS we have analyzed two spectral NMR datasets obtained from 60 human urine samples and 120 liver extracts, as well as an exhaustive spectral data simulation under a large number of parametric scenarios. According to our results, FOCUS methodology is an optimal data processing workflow for 1D-NMR analysis, and its accuracy and efficiency makes it suitable for the forthcoming large-scale metabolomic studies.

5 | Identification of Risk Loci for Crohn's Disease Phenotypes Using a Genome-Wide Association Study

Note: This chapter is an exact copy of the paper:
A. Alonso, E. Domènech, A. Julià, et al.. **Identification of Risk Loci for Crohn's Disease Phenotypes Using a Genome-wide Association Study**.
In Press Accepted Manuscript. DOI: 10.1053/j.gastro.2014.12.030.

Abstract

BACKGROUND & AIMS: Crohn's disease is a highly heterogeneous inflammatory bowel disease comprising multiple clinical phenotypes. Genome-wide association studies (GWASs) have associated a large number of loci with disease risk but have not associated any specific genetic variants with clinical phenotypes. We performed a GWAS of clinical phenotypes in Crohn's disease. **METHODS:** We genotyped 576,818 single nucleotide polymorphisms in a well-characterized cohort of 1090 Crohn's disease patients of European ancestry. We assessed their association with 17 phenotypes of Crohn's disease (based on disease location, disease behavior, disease course, age at onset, and extra-intestinal manifestations). A total of 57 markers with strong associations to Crohn's disease phenotypes ($P < 2 \cdot 10^{-4}$) were subsequently analyzed in an independent replication cohort of 1296 patients of European ancestry. **RESULTS:** We replicated the association of 4 loci with different Crohn's disease phenotypes. Variants in *MAGI1*, *CLCA2*, *2q24.1*, and *LY75* loci were associated with a complicated stricturing disease course ($P_{combined} = 2.01 \cdot 10^{-8}$), disease location ($P_{combined} = 1.3 \cdot 10^{-6}$), mild disease course ($P_{combined} = 5.94 \cdot 10^{-7}$), and erythema nodosum ($P_{combined} = 2.27 \cdot 10^{-6}$), respectively. **CONCLUSIONS:** In a GWAS, we associated 4 loci with clinical phenotypes of Crohn's disease. These findings indicate a genetic basis for the clinical heterogeneity observed for this inflammatory bowel disease.

5.1 Introduction

Crohn's disease (CD, OMIM(266600)) is a prevalent chronic inflammatory disease characterized by segmental and transmural inflammation of the gastrointestinal tract. Beyond clinical symptoms, persistent intestinal inflammatory activity leads to disease-related complications such as intestinal stenosis, fistulas, and abscesses. These complications are rarely managed medically and often require intestinal resection[6].

Although the precise etiology of CD remains unknown, it is commonly accepted that it is caused by the interaction of both genetic and environmental factors[209]. Recently, the genetic basis of CD risk has been exhaustively investigated by genome-wide association studies (GWAS). These studies have been highly successful, identifying 140 loci associated with CD susceptibility[210;211]. These loci have provided valuable insights into the biological processes involved in CD etiology, highlighting the importance of the dysregulation of the mucosal immune system and the loss of intestinal barrier integrity in the development of CD[211].

CD is known to be a highly heterogeneous disease showing different patterns of disease location, clinical behaviour, and complications[6;7]. From a clinical perspective, the study of the more severe disease phenotypes, those which require surgical intervention or anti-TNFα therapy, is of high importance. These phenotypes have shown a significant level of aggregation within individual families[8;9], suggesting that there is a genetic basis for disease heterogeneity. To date, several candidate gene association studies[10–14] have investigated the association of previously known CD risk loci to disease phenotypes, but only genetic variants in *NOD2* locus have been consistently validated[15;16].

In the present work, we have performed the first GWAS on CD clinically relevant phenotypes. For this objective we used a large multicenter cohort of CD patients[207] of Southern European ancestry. Using an independent cohort of CD patients of the same ancestry we performed a replication study to validate the most significant loci associated with CD clinical phenotypes.

5.2 Patients and methods

5.2.1 Patient Subjects

A total of 1,338 patients fulfilling Lennard-Jones diagnostic criteria for CD[212] were enrolled in the discovery phase (GWAS) of this study. Patient recruitment was performed between June 2007 and December 2010 at 15 Gastroenterology departments from different Spanish university hospitals belonging to the Immune-Mediated Inflammatory Disease Consortium (IMIDC)[207]. The IMIDC is a Spanish network of researchers investigating the genomic basis

of immune-mediated inflammatory diseases. A cohort of 1,493 healthy individuals of the same ancestry and previously described in Julià et al.[207], was used to control for the presence of genetic stratification and to evaluate the association to overall CD susceptibility of those loci associated with CD phenotypes. Both CD and control cohorts were Caucasian Europeans and all their four grandparents were born in Spain.

An independent cohort of 1,627 CD patients was used to replicate the most significant loci associated with the different CD phenotypes in the GWAS. The clinical selection criteria were the same as in the discovery cohort, with all patients fulfilling Lennard-Jones diagnostic criteria. All the patients included in the replication cohort were Caucasian Europeans and born in Spain. The patients were recruited both from the IMIDC and from the ENEIDA project[213] of the Spanish Working Group in IBD (GETECCU) repository.

An exhaustive record of epidemiological and clinical data was collected for each patient included in the study (Table 5.1). Informed consent was obtained from all participants, and protocols were reviewed and approved by local institutional review boards. This study was conducted in accordance with the Declaration of Helsinki principles.

5.2.2 Crohn's Disease Phenotypes

CD phenotypes for the GWAS were selected according to their clinical relevance and following the consensus criteria defined by the Montreal classification[7] (Table 5.2). Association was tested for the main phenotypic disease parameters: disease location (ileal (L1), colonic (L2), ileocolonic (L3), and upper disease (L4)), disease behaviour (inflammatory (B1), stricturing (B2), and penetrating (B3)), age at disease onset (early onset ($\leq$16 years) and late onset (>40 years)), and presence of perianal disease. We also tested for association the more frequently reported extraintestinal manifestations: erythema nodosum and peripheral arthropathy. When analyzing disease behaviour we restricted the control group to those B1 patients having a follow-up time >10 years (herein defined as $B1^+$ group). The same restriction was applied when analyzing perianal disease (i.e. the control group was limited to patients with no perianal disease and with >10 years of follow-up).

In addition to the phenotypes defined above, we also studied the association with additional disease outcomes related with disease severity and evolution. For this purpose we analyzed four different complicated disease course (CDC) outcomes: stricturing, penetrating, non-luminal (defined herein as stricturing or penetrating) and perianal disease. Importantly, only patients that had undergone surgery or had received anti-TNFα treatment were included in the CDC outcomes (Table 5.2). A mild disease course (MDC) phenotype was defined as the fulfillment of the following criteria: inflammatory behaviour, follow-up

Variable	Discovery cohort[1]	Replication cohort[1]	*P*-Value	OR (95%CI)
Sample size (after quality control)	1090	1296		
Females	551/1090 (50.55%)	658/1296 (50.77%)	0.93	1.01 (0.86-1.19)
Age at diagnosis (years)	26.95 [21.16-35.59]	29.12 [22.49-41.24]	$3.54 \cdot 10^{-8}$	1.02 (1.01-1.02)
Follow-up time (years)	15.48 [11.80-20.54]	10.65 [6.09-16.56]	$4.34 \cdot 10^{-43}$	0.92 (0.91-0.93)
Smokers at diagnosis time	465/1090 (42.66%)	466/1296 (35.96%)	$8.74 \cdot 10^{-4}$	0.75 (0.64-0.89)
Anti-TNF	417/1090 (38.26%)	467/1102 (42.38%)	0.05	1.19 (1.00-1.41)
Ileal location (L1)	242/833 (29.05%)	324/967 (33.51%)	0.05	1.23 (1.00-1.51)
Colonic location (L2)	169/833 (20.29%)	187/967 (19.34%)	0.64	0.94 (0.74-1.20)
Ileocolonic location (L3)	422/833 (50.66%)	456/967 (47.16%)	0.14	0.87 (0.72-1.05)
Upper disease (L4)	110/746 (14.75%)	112/864 (12.96%)	0.31	0.86 (0.64-1.16)
Behaviour B1	460/974 (47.23%)	728/1245 (58.47%)	$1.63 \cdot 10^{-7}$	1.57 (1.32-1.87)
Inflammatory behaviour ($B1^+$)[2]	389/970 (40.10%)	330/1244 (26.53%)	$1.60 \cdot 10^{-11}$	0.54 (0.45-0.65)
Stricturing behaviour (B2)	340/974 (34.91%)	286/1245 (22.97%)	$7.84 \cdot 10^{-10}$	0.56 (0.46-0.67)
Penetrating behaviour (B3)	265/974 (27.21%)	280/1245 (22.49%)	0.01	0.78 (0.64-0.95)
Perianal disease	361/1067 (33.83%)	330/1277 (25.84%)	$2.81 \cdot 10^{-5}$	0.68 (0.57-0.82)
Not perianal disease[2]	614/1061 (57.87%)	468/1276 (36.68%)	$1.27 \cdot 10^{-24}$	0.42 (0.36-0.50)
CDC-B2	261/650 (40.15%)	213/543 (39.23%)	0.77	0.96 (0.76-1.22)
CDC-B3	196/585 (33.50%)	173/503 (34.39%)	0.8	1.04 (0.80-1.35)
CDC-Non-Luminal	382/771 (49.55%)	344/674 (51.04%)	0.6	1.06 (0.86-1.31)
CDC-Perianal	143/757 (18.89%)	100/568 (17.61%)	0.57	0.92 (0.68-1.23)
MDC	173/1013 (17.08%)	113/1019 (11.09%)	$1.24 \cdot 10^{-4}$	0.61 (0.46-0.79)
Early disease onset	75/1067 (7.03%)	96/1287 (7.46%)	0.75	1.07 (0.77-1.48)
Late disease onset	184/1067 (17.24%)	343/1287 (26.65%)	$4.38 \cdot 10^{-8}$	1.74 (1.42-2.15)
Bowel resection	403/1090 (36.97%)	335/1102 (30.40%)	$1.14 \cdot 10^{-3}$	0.74 (0.62-0.89)
EIM Erythema nodosum	85/1090 (7.80%)	54/1102 (4.90%)	$6.42 \cdot 10^{-3}$	0.61 (0.42-0.88)
EIM Arthropathy	215/1090 (19.72%)	146/1102 (13.25%)	$5.32 \cdot 10^{-5}$	0.62 (0.49-0.79)
Family history[3]	101/1090 (9.27%)	84/1296 (6.48%)	0.01	0.68 (0.50-0.93)

[1] For categorical phenotypes n/N (%) where n is the number of patients displaying the related phenotype, N the total number of patients and % is the percentage of patients with the phenotype. For quantitative phenotypes (e.g. age at diagnosis) M [IQR] where M is the median value and IQR is the inter-quartile range.
[2] Control groups of perianal disease and behavior phenotypes require absence of the severe outcome and follow-up time > 10 years.
[3] ≥1 1st- or 2nd-degree relatives diagnosed with CD.

Table 5.1: Distribution and comparison of subphenotypes and clinical variables on the discovery and replication cohorts. Comparison of both cohorts was conducted using Fisher's exact test for categorical variables and logistic regression test for quantitative variables.

time >10 years, absence of perianal disease, and without the need of surgery or starting of anti-TNFα treatment.

After defining the most relevant disease phenotypes in CD we performed a total of 17 GWAS (Table 5.2).

5.2.3 Genotyping in Discovery and Replication Analysis

Genome-wide genotyping was performed using the Illumina Quad610 Beadchips (Illumina, USA) at the Centro Nacional de Genotipado (CeGen, Spain). Genotype calling was per-

Variable	Negative group[1]	N_{neg}[2]	Positive group[3]	N_{pos}[2]
Ileal involvement	L2	169/187	L1+L3	664/780
Colonic involvement	L1	242/324	L2+L3	591/643
Purely ileal vs purely colonic	L1	242/324	L2	169/187
Upper disease	Not L4	636/752	L4	110/112
Stricturing disease	$B1^+$	389/330	B2	340/286
Penetrating disease	$B1^+$	389/330	B3	265/280
Age at onset	Quantitative trait (1067/1285)			
Early onset	A2+A3	992/1191	A1	75/96
Late onset	A1+A2	883/944	A3	184/343
Perianal disease	Absence and Follow-up > 10 years	614/468	Presence	361/330
CDC-B2	$B1^+$	389/330	B2 and [Abdominal surgery or anti-TNFα therapy]	261/213
CDC-B3	$B1^+$	389/330	B3 and [Abdominal surgery or anti-TNFα therapy]	196/173
CDC-Non-Luminal	$B1^+$	389/330	[B2 or B3] and [Abdominal surgery or anti-TNFα therapy]	382/344
CDC-Perianal	Absence Perianal disease and Follow-up > 10 years	614/468	Perianal disease and [Perianal surgery or anti-TNFα therapy]	143/100
MDC	Not B1 or Perianal dis. or Anti-TNF TNFα therapy or Surgery	840/906	B1+and No surgery and No anti-TNFα therapy and No perianal disease	173/113
EIM Erythema Nodosum	Absence	1005/1048	Presence	85/54
EIM Arthropathy	Absence	875/956	Presence	215/146

[1] Group used as contrast in the analysis of the indicated variable.
[2] Number of samples in the negative and positive phenotype subgroups (Discovery/Validation).
[3] Positive group for the indicated variable.

Table 5.2: Clinical phenotypes studied in the GWAS analyses. This Table shows the phenotypes that were tested for in the GWAS discovery phase.

formed using the Illumina GenomeStudio software v2010.1 (Illumina, USA). A total of 576,818 markers and 1,338 samples were selected for quality control (QC) analysis (Supplementary Materials and Methods C.3). Potential population stratification was evaluated using the principal component analysis (EIGENSTRAT[214]; Supplementary Figure C.1). After QC analysis, a final dataset of 539,846 SNPs and 1,090 CD patients were available for the clinical subphenotype GWAS.

SNP imputation to further investigate associated loci in the GWASs was performed using SHAPEIT V2-644[215] and IMPUTE V2[216] software. We used the data from the European cohort from the 1,000 Genomes Project[217] as the reference panel (Supplementary Materials and Methods C.3).

The replication genotyping was performed at the HudsonAlpha Institute for Biotechnology

(Huntsville, Alabama, USA) on an independent cohort of 1,627 CD patients. Fifty-seven SNPs were selected for replication (Supplementary Materials and Methods C.3) and genotyped using the Illumina GoldenGate assay (Illumina, USA). All these SNPs had a call rate over 97.5%. Genotyping error rate of 0.32% was estimated using a subset of samples (i.e. 5% of the total) that were genotyped twice. After excluding samples with incomplete clinical data or missing rates $\geq$10%, a final set of 1,296 patients was available for the validation of the phenotypic associations identified in the GWAS stage.

5.2.4 Statistical Analysis

Descriptive statistics were computed to characterize relevant epidemiological and phenotypical variables in the studied cohorts (i.e. discovery and replication cohorts). The distributions of each variable in both cohorts were compared using Fisher's exact test and logistic regression test for qualitative and quantitative variables, respectively (Table 5.1). Phenotype co-occurrence was investigated by evaluating the association of CD phenotypes with respect to other variables of interest (Supplementary Material and Methods C.3). Using the final dataset of QC-filtered SNPs, the GWASs were performed using the allelic χ^2 and Wald tests implemented in PLINK software[218] for binary and quantitative variables, respectively. SNP selection for replication was performed using two criteria. First, the SNPs showing the strongest statistical significance ($P < 5 \cdot 10^{-6}$) and representing independent loci were automatically selected. Second, for those markers showing a moderate level of association (defined as $5 \cdot 10^{-6} < P < 5 \cdot 10^{-4}$), we performed an evaluation of the functional/regulatory impact of the region harboring the SNP (Supplementary Material and Methods C.3) and we selected those displaying biological evidence[219]. In order to discard the presence of confounding effects, a multivariate analysis including gender, smoking habit as well as the two first principal components of variation was performed on all the SNP-phenotype associations selected for replication. For the association analysis of the replication study we also used the allelic χ^2 and Wald tests. The criteria for replication included significance at the nominal level ($P < 0.05$) and same direction of effect than the observed in the GWAS. The combined analysis of GWAS and replication cohorts was performed by testing the association in the combined (GWAS and Replication) dataset. For the replicated SNP-Phenotype associations we performed a logistic regression model adjusting by gender, smoking habit, age at diagnosis and the phenotype's risk factors as identified in the phenotype co-occurrence analysis. In order to evaluate the potential regulatory role of each replicated loci, we used the expression quantitative trait loci (eQTL) browser (Pritchard Lab, *http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl*, University of Chicago, USA) as well as available regulatory data from the ENCODE project[220]. The NCBI GEO dataset GSE41269[221] was also used to evaluate intestinal tissue (i.e. ileal biopsies) specific eQTLs related to the SNPs with replicated phenotype associations (Supplementary Material and Methods C.3).

In addition to the GWAS, we performed a validation study of the previously reported SNP-phenotype associations. The validation was performed at the SNP level and at the locus level (Supplementary Material and Methods C.3). A total of 91 candidate SNP-Phenotype associations were identified and analyzed in the discovery cohort. Given the high relevance of *NOD2* locus in CD etiology, we also analyzed the association of the *NOD2* polymorphisms with all the clinically relevant phenotypes in CD.

In order to analyze the effects of associated SNPs in the time from diagnosis to the development of stricturing behaviour we performed a survival analysis using Kaplan-Meier estimator and Cox proportional hazard regression model (i.e. reported as hazard ratios (HR)). This analysis was performed using the R package Survival[222].

The remaining statistical analyses were conducted using R software for statistical computing (version 3.0.1)[223]. Association graphics with linkage disequilibrium patterns were generated using the R package snp.plotter[224].

5.3 Results

5.3.1 Phenotypic Characterization of the Studied Cohorts

The clinical and epidemiological characteristics of the discovery and replication patient cohorts are shown in Table 5.1. The GWAS cohort was characterized by a markedly higher follow-up time than the replication cohort (Median [Interquartile range]; 15.48 [11.80-20.54] vs 10.65 [6.09-16.56] years; $P = 4 \cdot 10^{-43}$). Consequently, there was also a statistically significant higher percentage of patients with severe clinical phenotypes like B2, B3, and perianal disease in the GWAS cohort compared to the replication cohort. However, when looking at the CDC phenotypes, the differences between both cohorts were no longer significant due to the normalization effect of using the follow-up time restriction. The results obtained in the phenotype co-occurrence analysis are shown in Supplementary Table 1 and Supplementary Figures C.2 to C.4.

5.3.2 Validation of Previously Associated Loci

From the total of 91 previously described genetic associations excluding *NOD2* locus, we successfully validated *ATG16L1*, *NCF4*, and *FOXO3* loci associations with inflammatory behaviour[13] ($P = 0.042$; OR 1.22; 95% CI 1.01-1.48), perianal disease[12] ($P = 0.046$; OR 1.21; 95% CI 1-1.46) and aggressive disease[14] ($P = 0.017$; OR 0.64; 95% CI 0.44-0.93), respectively (Table 5.3 and Supplementary Tables 2 and 3). In the gene-set analysis we also

replicated several significant phenotype associations which include *IRGM*[10], *NLRP1*[11], *TCF-4*[225], *MIF*[226], *SMAD3*[227], and *HNF4A*[228] loci (Table 5.3). Supplementary Figure C.5 shows the loci association *P*-Values of the validated associations.

We also replicated several previously reported associations of *NOD2* with different CD clinical phenotypes (Table 5.3 and Supplementary Table 4). The most significant association was between intronic SNP rs62029864 and bowel resection ($P = 4.5 \cdot 10^{-6}$; OR 1.58; 95% CI 1.30-1.93; Figure 5.1).

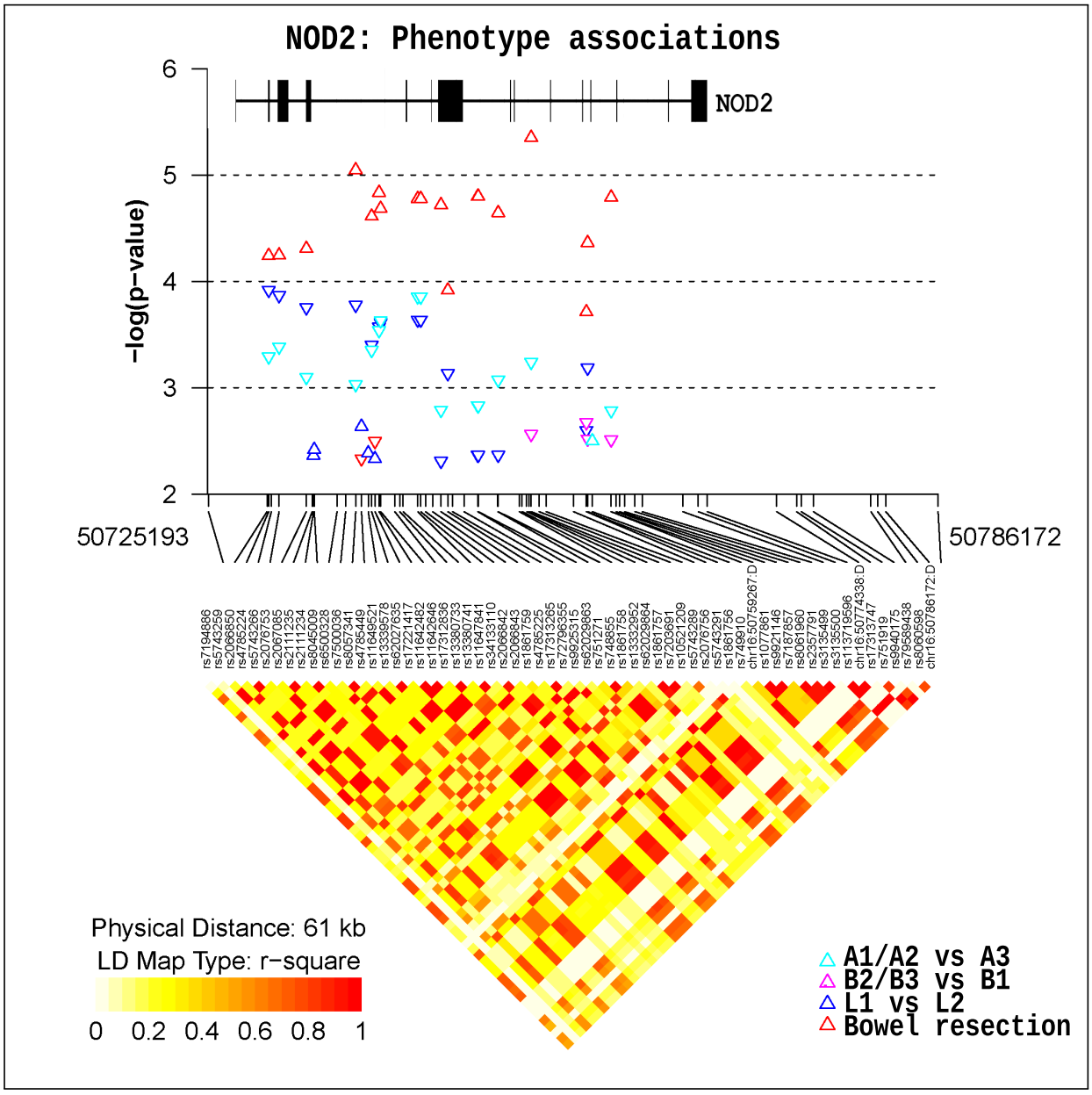


Figure 5.1: *Association between CD phenotypes and NOD2 imputed variants reaching significant values of association.* The upward triangles symbolize the *NOD2* variant confers risk for A3, B1, L2, or bowel resection. The downward triangles symbolize the risk is conferred for A1/A3, B2/B3, L1, and absence of bowel resection.

5.3.3 Phenotype GWAS and Replication Study

We performed 17 GWAS to identify new loci associated with clinically relevant phenotypes in CD (Table 5.2). Principal component analysis showed no genetic ancestry differences between the CD cohort and the control population of Southern European ancestry (Supplementary Figure C.1). At the phenotype level, we found no significant correlations between each of the analyzed clinical phenotypes and the estimated principal components of variation (Supplementary Figure C.6). Furthermore, no significant deviations from expectations were observed on quantile-quantile plots within the results of each analyzed phenotype, discarding the presence of systematic bias (Supplementary Figure C.7). SNP intensity cluster plots were visually inspected to rule out the presence of genotyping errors (Supplementary Figure C.8).

A total of 57 SNPs associated with clinical phenotypes in the GWAS phase (Supplementary Figures C.9 to C.11) were selected for replication. Importantly, only 5 of these associated SNPs mapped to a previously known susceptibility locus for CD (Supplementary Table 5).

Using an independent cohort of 1,296 CD patients, we validated the association of four SNP-phenotype associations ($P < 0.05$, same direction of effect as the GWAS) (Table 5.4 and Figure 5.2). The association of *MAGI1* locus on chromosome *3p14.1* with CDC-B2 phenotype was robustly validated in the replication cohort (rs11924265; $P = 3.04 \cdot 10^{-4}$; OR 2.27; 95% CI 1.41-3.68). Combining the association evidence from the GWAS and replication cohorts, this genetic association reached the genome-wide level of significance ($P = 2.01 \cdot 10^{-8}$; OR 2.27; 95% CI 1.68-3.09). *CLCA2* locus association with ileal involvement (rs2249296; $P = 0.033$; OR 0.72; 95% CI 0.52-0.99), *2q24.1* locus association with MDC (rs1520339; $P = 0.025$; OR 0.69; 95% CI 0.49-0.96) and *LY75* locus association with erythema nodosum (rs10929956; $P = 0.044$; OR 0.66; 95% CI 0.43-1.01) were also replicated at the nominal level in the independent validation cohort. All the replicated phenotype associations showed no significant variation when adjusted by potential confounding variables (Supplementary Figure C.12) and were not associated to CD risk.

5.3.4 *MAGI1* Association to Stricturing Behaviour

The SNP rs11924265, associated with CDC-B2, is located in the first intron of *MAGI1* gene. The healthy control population showed a MAF between the two patient groups ($MAF_{control} = 0.092$, $MAF_{B1+} = 0.063$ and $MAF_{CDC-B2} = 0.133$ in the discovery cohort). Consequently, no significant association was found when testing for association with overall CD susceptibility. The associated SNP is located in a 46.5 Kb haplotype block inside a *MAGI1* intron (from 65,748,962 to 65,795,463 pairbase in chromosome 3; Figure 5.2A and Supplementary Figure C.13). We found no evidence for regulatory function within this genomic region. In

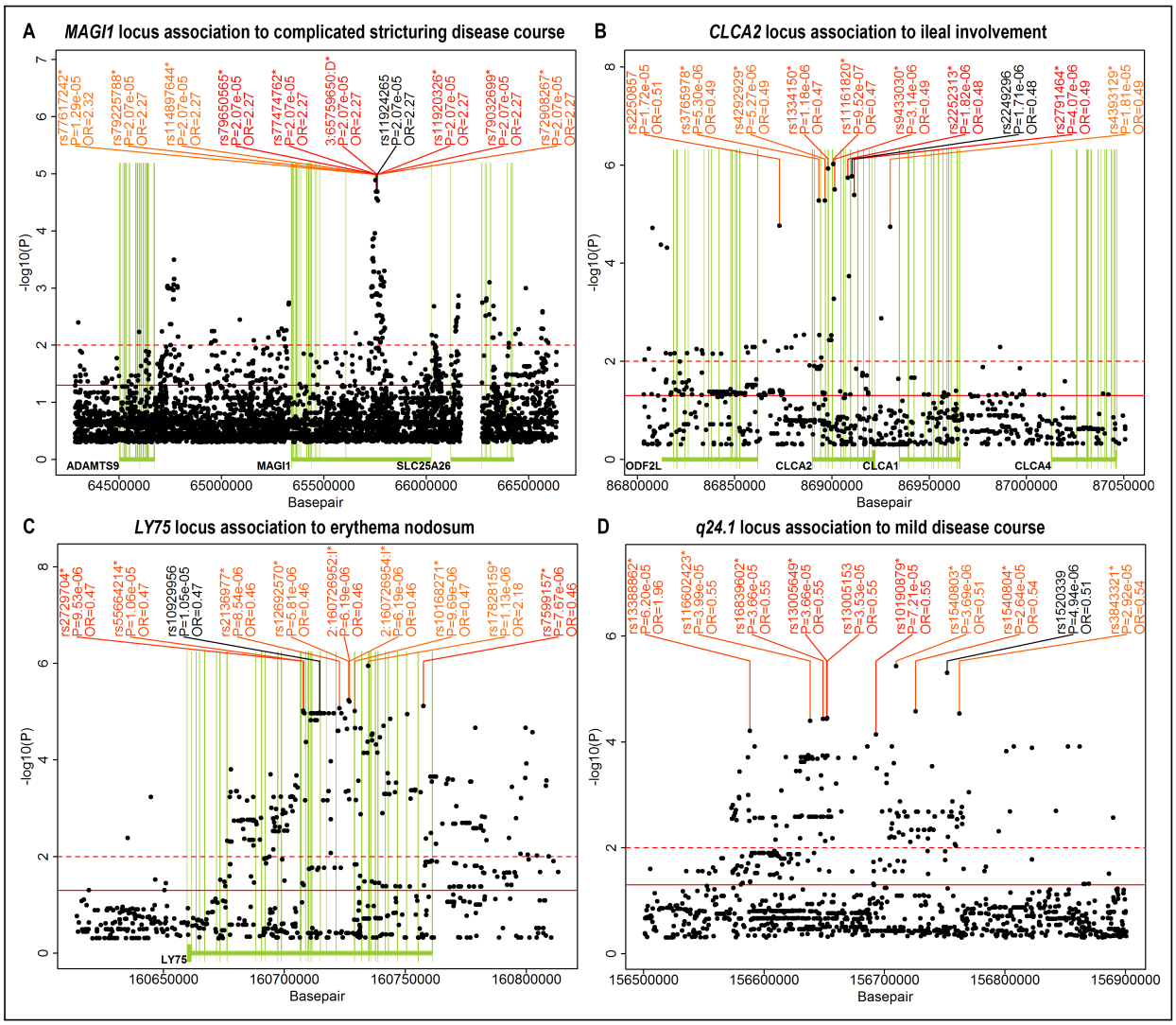


Figure 5.2: *Association statistics for the four SNP-Phenotype validated loci.* *MAGI1* (A), *CLCA2* (B), *LY75* (C) and *2q24.1* (D) loci associations with CDC-B2, ileal involvement, erythema nodosum, and MDC, respectively. SNPs annotated in black are the SNPs associated in the GWAS and validated in the replication study. Asterisks indicate imputed SNPs. Green lines outline gene exon locations. Continuous and dashed red lines indicate $-log_{10}$*P*-Values of 0.05 and 0.01, respectively.

order to evaluate the effect of the follow-up cut-off time on *MAGI1* association, we analyzed the change in effect size (i.e. OR) when increasing the follow-up time cut-off values for B1 patients to be included in the B1$^+$ (i.e. default >10 years) from 2 to 20 years. The results showed an increment in OR as the follow-up time threshold increased (Figure 5.3 A).

In addition to the association with CDC-B2 phenotype, we also found a strong association of SNP rs11924265 with B2 behaviour (Supplementary Figure C.13). The association between this SNP and B2 behaviour was nominally significant in the GWAS cohort ($P = 2.29 \cdot 10^{-4}$; OR, 1.99; 95% CI 1.35-2.95) and replicated in the validation cohort with a very similar effect size ($P = 6.66 \cdot 10^{-4}$; OR, 2.09; 95% CI 1.33-3.33). In the combined cohort the association with B2 behaviour was near to the genome-wide level of significance ($P = 5.34 \cdot 10^{-7}$; OR,

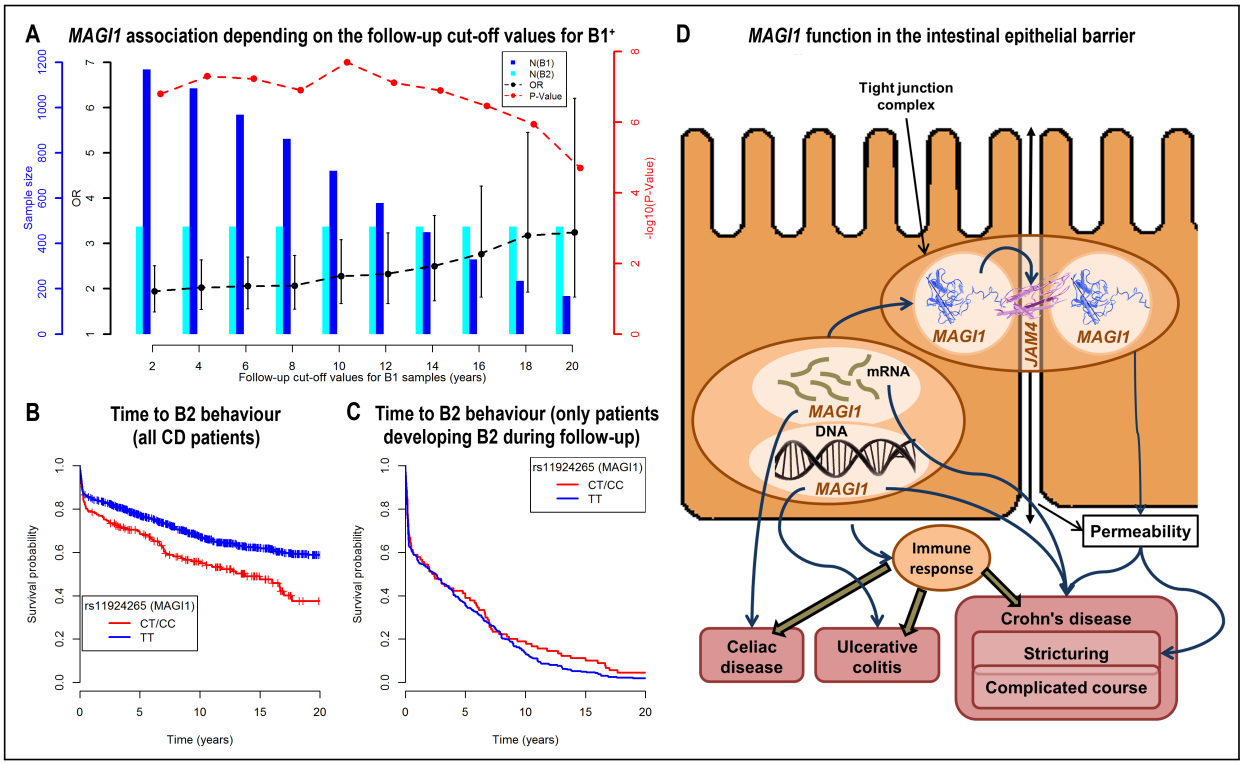


Figure 5.3: *Association results of MAGI1 locus and its role in epithelial barrier integrity.* (A) rs11924265 association to B2 depending on the follow-up cut-off value for $B1^+$ patients. As the cut-off value increases the number of available $B1^+$ patients decrease (i.e. dark blue bars) but the association effect increases (dashed black line). This may be due to the fact that the larger the follow-up cut-off value the smaller the number of B1 patients that will develop strictures in the future. (B) Kaplan-Meier curves of stricturing onset for risk allele carriers and non-carriers. (C) Kaplan-Meier curves restricted to CD patients that develop stricturing disease during follow-up. (D) Biological functions of MAGI1 in the intestinal epithelial barrier and previously reported associations with autoimmune diseases.

2.03; 95% CI 1.52-2.73). Survival analyses evaluating the time from diagnosis to the development of B2 behaviour also revealed significant differences between risk allele carriers and non-carriers ($P = 6 \cdot 10^{-5}$; HR, 1.64; 95% CI 1.30-2.08, Figure 5.3B). When restricting this analysis to patients that developed B2 behaviour during follow-up (Figure 5.3C), we found no significant differences between risk allele carriers and non-carriers indicating that the risk allele is related with a major B2 prevalence but it does not predispose to an early onset of this phenotype.

5.3.5 *CLCA2*, *LY75* and 2q24.1 Associations to CD phenotypes

The SNP rs2249296, associated with ileal involvement, is located in the 10th intron of *CLCA2* gene (Figure 5.2B). When compared to the control cohort and to patients with ileal involvement, patients with purely colonic involvement showed a significantly higher MAF ($MAF_{control} = 0.13$, $MAF_{L1/L3} = 0.12$ and $MAF_{L2} = 0.22$ in the discovery cohort). SNP rs2249296 lies in an 8.2 kb haplotype block that spans multiple exons of the *CLCA2* gene.

We found evidence of cis-eQTL of this SNP with the expression of *ODF2L* ($P = 2.5 \cdot 10^{-14}$), a gene located at 48Kb from the SNP[229]. Given that *NOD2* is also strongly associated with ileal involvement, we tested for a genetic interaction (i.e. epistasis) between both genes and the risk to develop ileal involvement. We did not find any evidence of interaction, suggesting that the two loci are independently associated with this trait.

The SNP rs10929956, associated with erythema nodosum, is located in an intronic region of *LY75* gene. Its minor allele (T) was found to be protective against the phenotype. The MAF in the health control cohort was similar to the MAF in the non-affected group of patients ($MAF_{control} = 0.43$, $MAF_{unaffected} = 0.46$, and $MAF_{affected} = 0.28$ in the discovery cohort). The associated SNP is located in an 86kb haplotype block (Figure 5.2C) which spans multiple *LY75* exons. Importantly, one of the most strongly correlated SNPs with rs10929956 is a nonsynonymous SNP (D807E) located in the 18th exon of *LY75* gene (rs3951216; $r^2 = 1.00$; $P_{GWAS} = 1.09 \cdot 10^{-5}$; Supplementary Table 6). We also identified a significant ileal tissue specific cis-eQTL ($P = 1.7 \cdot 10^{-4}$) for *LY75* gene with another strongly correlated SNP (rs10168271; $r^2 = 0.99$; $P_{GWAS} = 9.69 \cdot 10^{-6}$). For this analysis we used the genotyping and expression data on ileum samples (NCBI GEO dataset GSE41269; Supplementary Table 7 and Supplementary Figure C.14).

The association between *2q24.1* locus SNP rs1520339 and MDC achieved a near genome-wide significance in the combined dataset ($P_{combined} = 5.94 \cdot 10^{-7}$; $OR_{combined}$, 0.58; 95% CI 0.47-0.72). This SNP is located in a 132 kb haplotype block in a gene desert region in the long arm of chromosome 2 (Figure 5.2D). We found evidence of a trans-eQTL association with the expression of *PDE4D* gene in ileal tissue ($P = 1.4 \cdot 10^{-4}$; Supplementary Table 7 and Supplementary Figure C.14).

Gene ($SNP_{reported}$)	$OR_{reported}$	$P_{reported}$	Phenotype	P_{SNP} [1]	OR_{SNP} [1]	N_{SNP} [2]	$P_{GeneSet}$ [2]	SNP_{Pmin} [2]	P_{min} [3]
FOXO3 (rs12212067)	0.62 (0.52-0.73)	$2.09 \cdot 10^{-8}$	Agressive	0.0172*	0.64 (0.44-0.93)	20	0.1	rs17598747	0.012
ATG16L1 (rs2241880)	1.75 (1.22-2.53)	0.002	B1	0.0422*	1.22 (1.01-1.48)	29	0.16	rs2241876	0.018
NCF4 (rs4821544)	1.47 (1.08-1.99)	0.01	Perianal disease	0.0465*	1.21 (1.00-1.46)	36	0.49	rs760519	0.023
IRGM (rs13361189)	1.57 (1.18-2.09)	0.0024	L3	0.256	0.86 (0.67-1.11)	16	0.04**	rs6869426	0.003
NLRP1 (rs2301582)	1.35 (1.05-1.73)	0.02	B1	0.386	0.92 (0.76-1.11)	21	0.02**	rs11657333	0.003
NLRP1 (rs4790777)	1.33 (1.04-1.72)	0.03	B1	0.404	0.92 (0.76-1.12)	18	0.02**	rs11657333	0.003
TCF-4 (rs3814570)	1.38 (1.03-1.84)	0.028	L4	0.811	0.96 (0.69-1.34)	15	0.04**	rs7917983	0.008
NLRP1 (rs11651270)	1.35 (1.06-1.73)	0.02	B1	0.831	0.98 (0.81-1.18)	23	0.01**	rs11657333	0.003
MIF (rs755622)	0.22 (0.06-0.75)	0.009	L4	0.877	1.03 (0.68-1.56)	42	0.02**	rs5760071	0.002
SMAD3 (rs17293632)	4.88 (2.26-10.53)	*nd*	Bowel resection	nd	nd	34	0.03**	rs17293443	0.003
HNF4A (rs1884613)	1.38 (1.18-1.63)	0.0001	Early onset	0.68	1.09 (0.71-1.68)	23	0.07	rs2425640	0.001***
NOD2[3]	1.73 (1.35-2.24)	$2.3 \cdot 10^{-5}$	Bowel resection	nd	nd	133	$1.68 \cdot 10^{-3}$**	rs62029864	$4.45 \cdot 10^{-6}$***
NOD2 (-)	-	-	Late onset	nd	nd	133	0.02**	rs13380733	$1.40 \cdot 10^{-4}$***
NOD2[3]	0.49 (0.30-0.81)	$4.0 \cdot 10^{-4}$	L1 vs L2	nd	nd	133	$4.92 \cdot 10^{-3}$**	rs2066850	$1.21 \cdot 10^{-4}$***
NOD2[3]	1.90 (1.46-2.47)	$2.0 \cdot 10^{-6}$	Ileum involv.	nd	nd	133	0.02**	rs5743266	$4.19 \cdot 10^{-4}$

[1] SNP-level validation. P_{SNP} and OR_{SNP} refer to the association statistics obtained at the tagSNP of the previously reported SNP.
[2] Locus-level validation. The number of SNPs included in the locus analysis (N_{SNP}), the gene-set test association *P*-Value ($P_{GeneSet}$), the most significantly associated SNP (SNP_{Pmin}) and its corresponding *P*-Value (P_{min}) are reported.
[3] When analyzing *NOD2* locus, SNP8 (rs2066844), SNP12 (rs2066845) and SNP13 (rs2066847) are usually condensed in one variable that accounts for the number of risk alleles in these three SNPs.
nd: No data available; *: Validated at the SNP level; **: Validated at the locus level in the gene-set analysis; ***: Validated at the locus level with Bonferroni correction.

Table 5.3: Previous SNP-phenotype associations replicated in our study. This Table shows the previously reported SNP-phenotype associations replicated in our discovery cohort. Significant *NOD2* associations for the analyzed phenotypes are reported at the locus level. Further details are given in Supplementary Tables 3 and 4

Phenotype	SNP[1]	Position[2]	Gene	MAF_{GWAS}[3]	P_{GWAS}	OR_{GWAS}(95CI)	MAF_{REP}[4]	P_{REP}	OR_{REP}(95CI)	P_{COM}	P_{CD}[5]
CDC-B2	rs11924265-C	3:65760342	*MAGI1* (intron)	0.06/0.13	$1.9 \cdot 10^{-5}$	2.27 (1.52-3.42)	0.05/0.11	0.0003	2.27 (1.41-3.68)	$2.0 \cdot 10^{-8}$	0.17
Eryth. Nodosum	rs10929956-T	2:160714615	*LY75* (intron)	0.46/0.28	$1.1 \cdot 10^{-5}$	0.47 (0.32-0.67)	0.46/0.36	0.0437	0.66 (0.43-1.01)	$2.3 \cdot 10^{-6}$	0.19
Ileal involvement	rs2249296-T	1:86910264	*CLCA2* (intron)	0.22/0.12	$1.7 \cdot 10^{-6}$	0.48 (0.35-0.66)	0.17/0.13	0.0331	0.72 (0.52-0.99)	$1.3 \cdot 10^{-6}$	0.53
MDC	rs1520339-T	2:156751921	*2q24.1*	0.31/0.19	$4.9 \cdot 10^{-6}$	0.51 (0.38-0.69)	0.31/0.24	0.025	0.69 (0.49-0.96)	$5.9 \cdot 10^{-7}$	0.73
B2	rs11924265-C	3:65760342	*MAGI1* (intron)	0.06/0.12	$2.3 \cdot 10^{-4}$	1.99 (1.35-2.95)	0.05/0.11	0.0007	2.09 (1.33-3.33)	$5.3 \cdot 10^{-7}$	0.17

[1] SNP-minor allele.
[2] Chromosome:basepair.
[3] Minor allele frequency on negative and positive groups as defined in Table 2.
[4] SNP association *P*-Value when comparing all CD patients against the healthy control cohort.
GWAS: GWAS cohort; REP: Replication cohort; COM: Combined cohort.

Table 5.4: SNP-Phenotype associations validated in the replication cohort. Association results at the GWAS and replication analyses for the four validated SNP-Phenotype associations

5.4 Discussion

In the present study we report, for the first time, the results of GWAS on clinically relevant phenotypes in CD. A total of seventeen CD phenotypes of clinical importance were analyzed which included disease behaviour, disease location, complicated disease course, age at onset, and other complications like perianal disease and extraintestinal manifestations. Using an independent cohort of patients, we have validated the strong GWAS association of *MAGI1* with stricturing behaviour and complicated stricturing disease course. This last phenotype association reached the GWAS level of significance in the combined cohort. *CLCA2*, *LY75*, and *2q24.1* loci associations with ileal involvement, erythema nodosum, and MDC were validated at the nominal level. Finally, we also report the first confirmatory evidence of several phenotype associations reported in previous candidate-gene studies.

MAGI1 gene encodes for the membrane-associated guanylate kinase, WW and PDZ domain-containing protein 1[230]. Using survival analyses we have found that this locus predisposes to stricturing behaviour but does not accelerate its onset. When increasing the follow-up time threshold for the control group (B1), the association effect increased accordingly due to the reduced number of misclassified patients in this group. This protein has been shown to be highly involved in the tight junction of intestinal epithelial cells through the interaction with JAM4, a junctional adhesion transmembrane molecule[231–234]. This interaction between transmembrane proteins (i.e. JAM4) and scaffolding proteins (i.e. MAGI1) is pivotal for the barrier function of epithelial tight junctions ensuring a highly selective barrier permeability[233;235] (Figure 5.3D). This barrier is a complex cell monolayer that separates the intestinal lumen from the lamina propria[236] and acts as a barrier against commensal bacteria and foreign antigens. Consequently, the disruption of its homeostasis[236–238] can have dramatic effects on the mucosal integrity, which has been shown to contribute to the development of CD and ulcerative colitis (UC)[236;237;239–241]. Importantly, an increased intestinal permeability has been described in patients with active CD as well as in their first-degree relatives[239;242–244]. This evidence strongly suggests that increased intestinal permeability is not only a consequence of intestinal inflammation occurring in CD patients, but it is rather a risk factor for CD. Previous studies[242] have shown a significant increase in intestinal permeability in patients with stricturing disease, while no significant differences were found in patients with inflammatory or penetrating disease. This is consistent with the reported association between *MAGI1* and complicated stricturing disease and could contribute to an overwhelming immune response[245;246] and to the subsequent transmural inflammation of the gastrointestinal tract (Figure 5.3D).

MAGI1 associations to immune-mediated gastrointestinal diseases including CD, UC, and celiac disease have been previously reported[247–250]. *MAGI1* SNP rs924022, at 54Kb of the

SNP associated to CDC-B2 in our study, has been previously associated to medically refractory UC, a severe progressive phenotype that requires colectomy[248]. Also, very recently, variation at *MAGI1* locus has been associated with CD susceptibility in a candidate gene study[250]. Of relevance, the low effect size reported for this locus in CD risk is likely to be the result of combining CD patients with different behaviour phenotypes. In our patient cohort we show that, when merging B1 and B2 patients, the joint risk allele frequency becomes closer to that of the control cohort, therefore leading to a non-significant association with disease risk. This result strengthens our hypothesis that there is a genetic basis for disease heterogeneity that is independent from overall disease susceptibility.

CLCA2 (chloride channel accessory 2) codifies for a protein that belongs to the calcium sensitive chloride conductance protein family. High expression levels of this gene have been previously reported[251] as a characteristic feature of the follicle-associated epithelium and M cells in mice. This specialized cell type of the intestinal epithelium is known to play an important role in the induction of immune responses against mucosal antigens through the antigen transcytosis mechanism[252]. *CLCA2* gene expression has been also implicated in colorectal cancer[253] and mouse models of UC[254]. Of importance, SNP associated with ileal involvement seems to have strong eQTL evidence with *ODF2L* (outer dense fiber of sperm tails 2-like) gene[229]. Further studies are required to characterize the specific role of this genetic variation in the development of ileal involvement.

SNP rs1520339 in the *2q24.1* locus has been associated to MDC. This marker lies in a large intergenic region for which there is yet no functional regulatory evidence. Consequently, functional studies analyzing relevant tissues in CD pathology (i.e. colon and ileum) are needed to provide more insights on the biological mechanisms influenced by variation at this locus. Nonetheless, using recent genotyping and genome-wide expression data on intestinal samples[221], we have identified a specific trans-eQTL of this locus with the phosphodiesterase 4D (*PDE4D*) gene. In this association, the allele associated with MDC is associated with a lower expression of the *PDE4D* gene. Of relevance, *PDE4D* has been associated with the proinflammatory activity in several autoimmune diseases and *PDE4D* inhibitors are actually being used for the management of Psoriatic Arthritis and Psoriasis[255]. At present, the inhibitors of the PDE4 family of genes are also being studied for treating inflammatory bowel diseases[256]. Future studies determining the biological mechanisms underlying the regulatory evidence between *2q24.1* locus and this proinflammatory enzyme are therefore warranted.

Finally, variation at *LY75* (lymphocyte antigen 75) locus has been found to be associated with the risk to develop erythema nodosum. There is evidence that SNPs in high LD with the replicated SNP have strong functional effects in ileal tissue (i.e. non-synonymous SNP rs3951216

and rs10168271 with strong cis-eQTL regulation of *LY75* gene expression). *LY75* gene encodes for a C-type lectin receptor that participates in the immune system response[257]. *LY75* protein has been shown to mediate antigen uptake and presentation, and is mainly expressed in skin and in plasmacytoid dendritic cells[258;259]. Importantly, erythema nodosum is characterized by the inflammation of the subcutaneous fat cells. Additional studies aimed at characterizing the impact of this genetic variation into protein functionality and immune cell activity will be required to determine the implication of *LY75* in the development of erythema nodosum in CD.

In conclusion, we have performed the first GWAS for clinically relevant traits in CD and we have found new risk loci for four different phenotypes. Importantly, these new loci have not been previously described as risk factors for CD. Our results therefore demonstrate the existence of a genetic component for disease heterogeneity that is independent of the genetic variation associated with the susceptibility to CD. Further functional studies of these new loci will provide a better understanding of the biological mechanisms that are involved in the development of these relevant clinical phenotypes in CD.

6 | Discussion

The methodological tools that have been developed in the course of this thesis have demonstrated a significant improvement in the performance of the typical processing frameworks of high-throughput genomics and metabolomics studies. All the algorithms have been tested exhaustively and their accuracies have been evaluated against the state-of-the-art methods using reference datasets. In addition to the methodological contributions, the GWAS of CD phenotypes has reported very relevant results in the genetics of CD that provide new clues for understanding the clinical heterogeneity of this disease.

6.1 GStream: High-throughput SNP and CNV genotyping

GStream solves two major methodological needs of the GWASs analytical workflow. First, there is a lack of a bioinformatics tool that efficiently integrates both SNP and CNV genotyping algorithms. This integration will be crucial to streamline GWAS analyses. Second, most of the commonly used methods for CNV genotyping at the genome-wide scale do not use the powerful information generated by the simultaneous analysis of multiple samples and, instead, they are based on independent, per-sample analysis. While the latter approach has proven to work well for large genomic CNVs, it clearly fails to identify and genotype small CNVs, therefore leading to high false negative rates. Small CNVs are highly relevant when studying genome variation since their population frequency is higher than large CNVs, and they can span important functional regions such as exons, introns and regulatory elements[260].

GStream, the method developed in this PhD thesis, efficiently uses the information generated by the SNP genotyping stage and by the analysis of multiple samples to significantly increase the genotyping sensitivity and accuracy in Illumina genotyping microarrays. We demonstrate that this new method clearly outperforms previous CNV genotyping methodologies. Additionally, GStream integrates CNV genotyping with SNP genotyping and has been designed to be computationally efficient when performing GWASs. GStream has been implemented in C++ as an open source tool (*www.urr.cat/GStream*).

The performance of the genotyping algorithms has been assessed using well characterized reference data generated by the HapMap project[51]. Although not the main objective of this tool, GStream has outperformed previous SNP genotyping methods, providing a superior call rate. When using the most recent Illumina microarray, SNP genotyping with GStream has allowed to include the complete genotype of >2,000 additional SNPs on average. GStream obtained the best global genotyping accuracy in all the microarray platforms tested (i.e. Human610-Quad, Human660W-Quad, Human1M-Duo, and HumanOmni1-Quad; Table 3.3 and Figure A.3).

GStream provides a major improvement in CNV genotyping compared to previous state-of-the art methodologies (i.e. PennCNV[76], QuantiSNP[77] and CNstream[148]). The performance of CNV genotyping has been exhaustively evaluated using data from three reference CNV characterization studies[24;142;143], and from the 1KGP project[53]. The power of GStream to detect and correctly genotype CNVs was found to be significantly higher than the previous state-of-the-art methods. GStream has also demonstrated high genotyping efficiency in cases where the other methods have relevant limitations like, for example, in those CNVs that span only a few number of probes, or when the intensity distributions corresponding to the different copy number states show partial overlapping. The multi-component intensity distribution model implemented in GStream (Figure 3.2) allows researchers to perform a comprehensive scan of the genome for additional CNVs, broadening their detection range to shorter, population-specific and/or previously uncharacterized CNVs (Figure 3.5).

In order to provide additional evidence of the superior performance of GStream, we have also identified several CNVs in strong linkage disequilibrium (LD) with SNPs associated to different traits in previous GWAS. These CNVs include several known risk loci such as *IRGM1* and *LCE3B/LCE3C* deletions, that have been associated to Crohn's disease and psoriasis susceptibility, respectively[154;155].Importantly, new strong CNV candidates have also been detected. For example, using GStream a previously uncharacterized deletion spanning two exons of *SLC2A9* gene was identified; this CNV is in high LD with SNPs previously associated to uric acid concentration[186]. In these cases, the highly correlated CNV could be a powerful causal candidate of the observed genetic association. Therefore, these results could reveal important insights into the causality of these trait associations, particularly in those studies when the reported associated SNPs are located in genomic regions with an unknown functional impact. A complete list of the newly identified CNVs has been made publicly accessible at *www.urr.cat/GStream/SNP_CNV*.

6.2 FOCUS: 1D-NMR processing workflow for high-throughput metabolomics studies

The second work included in this thesis has been the development and implementation of FOCUS, a processing workflow for high-throughput NMR-based metabolomics studies. FOCUS has been implemented using Matlab®as an open-source tool (*www.urr.cat/FOCUS*). FOCUS has been developed in order to solve three main challenges in the data processing workflow of NMR-based metabolomics studies:

- To provide a software tool that efficiently integrates all different algorithms associated with NMR spectra processing.
- To minimize the impact of technical variability on downstream statistical analyses.
- To significantly improve automatic metabolite identification in NMR processed data by using all the information available in the reference metabolite spectral databases.

The results provided by FOCUS have demonstrated to efficiently solve these three main challenges and to provide more accurate results than the previous methodologies. The alignment algorithm that has been developed, RUNAS (Recursive UNreferenced Alignment of Spectra), performs spectral alignment using the cross-correlation function between spectra as the alignment maximization function (Figure 4.1). In contrast to most of the currently available methodologies, RUNAS does not rely on the definition of a reference spectrum. The alignment procedure applies a spectral transformation that enhances peak shapes and reduces the alignment bias produced by the presence of peaks with highly unpaired intensities.

The performance of RUNAS has been assessed using simulated and real data. Using synthetic data we demonstrate that FOCUS significantly reduces the correlation matrix distance between the true aligned spectra and the spectra aligned by the algorithm compared to previous methodologies (Figure 4.4 and Table 4.1). Consequently, FOCUS provides the best alignment performance improving the subsequent peak detection stage and reducing the alignment-associated biases. The results of the simulation analysis also show the robustness of RUNAS against high degrees of spectral unalignment. When evaluating spectral alignment performance of FOCUS using a real human NMR dataset, we show that FOCUS is clearly superior to other state-of-the art methodologies. FOCUS, Icoshift[122] and COW[117], respectively, showed average spectral correlation improvements of 53.5%, 39.4%, and 35.6%

with respect to the unaligned spectral data set (Figure 4.5). Reducing the number of analyzed samples produced only a slight reduction in alignment performance in all three methods, with consistently better results with the FOCUS aligned data set. Consequently, the performance superiority of FOCUS is practically independent of the sample size of the study.

In order to evaluate the metabolite identification algorithm we used FOCUS on two real spectral datasets. The first dataset consisted in NMR spectra of 60 human urine samples; in this data, FOCUS was able to correctly identify 22 different metabolites that were manually verified, demonstrating the good performance of the identification procedure implemented in this algorithm. The metabolite identification procedure in the second dataset (i.e. 120 liver extract samples) correctly identified 20 metabolites, also confirming the good performance of the identification procedure for tissue extract samples.

Finally, the peak detection method implemented in FOCUS has also demonstrated to have a highly significant accuracy. This is due to the use of a frequency threshold that takes into account that true signal peak positions are correlated, while noise peaks are uncorrelated. Consequently, noise peaks will rarely exceed the frequency threshold at a given spectral position. The results show that FOCUS provides a superior dynamic range (i.e. 1:1000) compared to other NMR data processing methods of high-throughput metabolomics studies. This superior sensitivity, allows FOCUS to significantly increase the power of metabolomic studies to detect relevant metabolite associations.

Overall, FOCUS represents a major improvement in the analytical framework for high-throughput metabolomics studies. The integration of the different processing stages in a single tool and the improvements in spectral alignment, peak detection and metabolite identification algorithms allow a fast and reliable data processing stage for large sample size studies.

6.3 GWAS analysis of Crohn's disease phenotypes

CD is a prevalent and highly heterogeneous inflammatory bowel disease that comprises multiple clinical phenotypes. The most severe CD phenotypes are of high importance since they can lead to complicated disease progression requiring surgical intervention and involving a drastic reduction in the patients' quality of life. CD phenotypes have shown to have a significant level of aggregation within families[8;9], thereby indicating the presence of a genetic risk basis. Although previous GWASs have identified loci associated to disease risk, the association of specific genetic variants with clinical phenotypes has not been yet studied at the genome-wide scale.

During this doctoral thesis, the first GWAS to identify the genetic basis of phenotypic heterogeneity in CD has been performed. To date, several candidate gene studies have analyzed the association of previously known CD risk loci with the susceptibility to develop specific CD phenotypes (i.e. *NOD2*, *FOXO3* and *IRGM*). The genomic study included in this thesis has analyzed the whole genome variability in a large cohort of patients and has identified new risk loci for different CD phenotypes. Importantly, these new risk loci are not associated to disease risk, supporting the existence of an independent genetic component specific for disease heterogeneity.

Using a two-stage GWAS design, 17 different clinical phenotypes of CD have been analyzed. In this study, we have identified and subsequently replicated four loci associated to CD phenotypes: *MAGI1*, *CLCA2*, *LY75*, and *2q24.1* loci have been associated to stricturing behaviour, ileal involvement, erythema nodosum, and mild disease course, respectively.

The association of *MAGI1* to stricturing behaviour reached genome-wide significance (i.e. $P < 5 \cdot 10^{-8}$) when combining the discovery and validation cohorts. The protein coded by this gene is a scaffolding protein that mediates cell to cell adhesion of intestinal epithelial cells[231–234]. Importantly, genetic variation in this locus has been previously associated to severe phenotypes in autoimmune diseases of the digestive system ulcerative colitis and celiac disease[248;249]. Together, its association to multiple autoimmune bowel diseases and its regulation of cell adhesion in the intestinal epithelial tissue indicates that *MAGI1* is a powerful biological candidate to be involved in the physiopathology of these diseases.

Disease location is one of the major classification criteria for CD, where the disease can affect the colon, the ileum or both of them[7]. Depending on disease location the prognostic and the therapeutic strategy can differ considerably. The association of this phenotype with genetic variants in the *CLCA2* gene is of high relevance. High expression levels of this gene in follicle-associated epithelium and M cells have been reported in previous studies[251]. These two cell types are found in the intestinal epithelium and are known to play an important role in the immune response against mucosal antigens. *CLCA2* has also been implicated in colorectal cancer[253] and mouse models of ulcerative colitis[254] confirming its implication in bowel disorders. Together, all these evidences reveal the importance of this gene in the immune response and common disorders of ileal tissue.

Erythema nodosum is an extraintestinal manifestation of CD characterized by the inflammation of the fat cells under the skin. There is evidence that SNPs in high linkage disequilibrium with the associated SNP in the *LY75* gene could have strong functional effects in ileal tissue (i.e. eQTLs and non-synonymous mutations)[221]. *LY75* encodes for a C-type lectin receptor that participates in the immune system response[257]. This protein has been shown to mediate antigen uptake and presentation, and is expressed mainly in skin and in

plasmacytoid dendritic cells[258;259]. Importantly, these cells where the resulting protein has a clear biological function are present in the main affected tissue of erythema nodosum.

7 | Conclusions

Genomics

- Microarray-based technologies have allowed the parallel genotyping of hundreds of thousands of genetic variants and the analysis of large sample collections at an increasingly affordable cost. Illumina microarrays have been the most widely used genotyping platform to perform GWAS and identify new risk variants in common diseases. However, there is a lack of a software tools that provide accurate methods for CNV genotyping in this type genotyping platform. The bioinformatics tool developed during this thesis, GStream, provides an unprecedented accuracy for both SNP and CNV genotyping in Illumina platform microarrays. GStream has demonstrated to outperform state-of-the-art used methods both in terms of genotyping call rate and accuracy. Importantly, the CNV algorithm has shown to improve the CNV detection sensitivity, providing a major coverage of genetic variation than the commonly used algorithms. Consequently GStream provides an unprecedented power to identify relevant genetic associations in genome-wide SNP and CNV association analyses.

- In the present thesis work we have performed the first GWAS for clinically relevant traits in CD. We have identified new risk loci for 4 different CD severity phenotypes. Importantly, these new loci are not associated to disease susceptibility, thereby confirming the hypothesis that there is an independent genetic component that influences disease heterogeneity in CD.

- GStream has been actually integrated into the analysis pipeline of the IMID-Consortium GWA studies. In one of these studies, GStream has recently allowed the identification of an intergenic deletion between *ADAMTS9* and *MAGI1* genes in chromosome 3 associated with psoriatic arthritis risk[261]. Importantly, this new CNV is one of the first genetic risk factors specifically associated to the joint autoimmune component in Psoriatic Arthritis.

Metabolomics

- The results presented in this thesis show that FOCUS NMR analysis software addresses the main problems that affect the processing of high-throughput NMR metabolomic data. FOCUS provides an integrated workflow that performs all the necessary processing steps required to obtain a set of spectral measurements ready for the usual chemometric analyses. Importantly, the algorithm for spectral alignment (RUNAS) implemented in FOCUS, has demonstrated the best performance when processing moderate to highly unaligned spectral data sets. This improvement is due to the use of an innovative approach where no reference spectra are needed. The implemented algorithm is an iterative alignment method based on the spectral correlation of each pair of spectral samples and that uses fast Fourier transform to speed up computation. Furthermore, the peak detection method implemented in FOCUS also avoids the bias produced by outlier samples since it is based on the peak-sample frequency. Finally, FOCUS also provides a new and efficient method for metabolite identification, facilitating this time-consuming task to spectroscopists that previously had to perform this task manually.

- The usefulness of FOCUS has been demonstrated analyzing two spectral NMR data sets obtained from 60 human urine samples and 120 liver extracts. We have also used simulated spectral datasets under a large number of parametric scenarios for a better assessment of spectral alignment accuracies. The obtained results show that FOCUS methodology is an optimal data processing workflow for 1D-NMR data analysis. The accuracy and efficiency of this tool makes it suitable for the forthcoming large-scale metabolomic studies that will include very large numbers of biological samples.

- FOCUS is currently being applied in a high-throughput metabolomics study for the identification and validation of diagnostic and activity biomarkers in IMIDs. This study has analyzed the urine metabolome within two independent cohorts of >2,500 individuals including healthy controls and IMID patients. This study has identified and validated significant differences in urine metabolite levels between the IMID patients and the healthy controls. The disease activity analysis has also identified significant associations between metabolite concentrations and the disease activity scores at sample collection time. The results of this study have been currently submitted for publication (Alonso et al. "*Urine metabolome profiling in immune-mediated inflammatory diseases*").

8 | Publications

8.1 Research Papers in Indexed Journals

- Arnald Alonso, Sara Marsal, Raul Tortosa, Oriol Canela-Xandri and Antonio Julià. **GStream: improving SNP and CNV coverage on genome-wide association studies.** PloS one 8, no. 7 (2013): e68822.
  PMID: 23844243

- Arnald Alonso, Miguel A. Rodríguez, Maria Vinaixa, Raul Tortosa, Xavier Correig, Antonio Julià and Sara Marsal. **Focus: a robust workflow for one-dimensional NMR spectral analysis.** Analytical chemistry 86, no. 2 (2013): 1160-1169.
  PMID: 24354303

- Arnald Alonso, Eugeni Domènech, Antonio Julià, Julián Panés, Valle García-Sánchez, et al. **Identification of Risk Loci for Crohn's Disease Phenotypes Using a Genome-wide Association Study.** Gastroenterology (2014).
  PMID: 25557950

8.2 Review Papers in Indexed Journals

- Antonio Julià, Arnald Alonso and Sara Marsal. **Metabolomics in rheumatic diseases.** International Journal of Clinical Rheumatology 9, no. 4 (2014): 353-369.
  DOI: 10.2217/ijr.14.25

- Arnald Alonso, Sara Marsal and Antonio Julià. **Analytical methods in untargeted metabolomics: state of the art in 2015.** Front. Bioeng. Biotechnol. 3, no. 23 (2015).
  DOI: 10.3389/fbioe.2015.00023

8.3 Seminar and conference talks

- **Identification and validation of diagnostic and activity urinary metabolomic biomarkers in immune-mediated inflammatory diseases.** Annual European Congress of Rheumatology (EULAR), Rome, Italy, 2015.
- **Identificación de biomarcadores metabolómicos de diagnóstico y actividad en enfermedades inflamatorias mediadas por inmunidad.** Annual Congress of the Spanish Society of Rheumatology, Santiago de Compostela, Spain, 2014.
- **Identification of disease diagnostic and disease activity metabolomic biomarkers in immune-mediated inflammatory diseases.** Annual European Congress of Rheumatology (EULAR), Paris, France, 2014.
- **Example of biomarker discovery in inflammatory diseases by NMR metabolomics.** Summer course "Metabolomics an indispensable tool for research in life sciences", Rovira i Virgili University, Tarragona, Spain, 2014.
- **ASTREAM and FOCUS: Two powerful methodologies for mass spectrometry and nuclear magnetic resonance data analysis.** 7th Scientific Conference Vall d'Hebron Research Institute, Barcelona, Spain, 2013.
- **GStream: A fast and highly accurate tool for SNP and CNV genotyping with Illumina microarrays.** 6th Scientific Conference Vall d'Hebron Research Institute, Barcelona, Spain, 2012.

8.4 Poster presentations

- **Identificación y validación de biomarcadores metabolómicos urinarios de diagnóstico y actividad en enfermedades inflamatorias mediadas por inmunidad mediante resonancia magnética nuclear.** Annual Congress of the Spanish Society of Rheumatology, Sevilla, Spain, 2015.
- **ASTREAM and FOCUS: Two powerful methodologies for metabolomic data analysis in high-throughput studies.** 1st Scientific conference on Bioinformatics and computational biology, Barcelona, Spain, 2013.
- **CNStream2: a fast and highly accurate tool for SNP and CNV genotyping with Illumina microarrays.** 16th Annual International Conference on Research in computational Molecular Biology, Barcelona, Spain, 2012.

Bibliography

[1] Annabel Kuek, Brian L Hazleman, and Andrew JK Ostor. Immune-mediated inflammatory diseases (IMIDs) and biologic therapy: a medical revolution. *Postgraduate medical journal*, 83(978):251–260, 2007.

[2] A Boonen and W Mau. The economic burden of disease: comparison between rheumatoid arthritis and ankylosing spondylitis. *Clinical & Experimental Rheumatology*, 27(4):S112, 2009.

[3] Arnald Alonso, Sara Marsal, Raül Tortosa, Oriol Canela-Xandri, and Antonio Julià. GStream: Improving SNP and CNV Coverage on Genome-Wide Association Studies. *PloS one*, 8(7):e68822, 2013.

[4] Arnald Alonso, Miguel A. Rodríguez, Maria Vinaixa, Raul Tortosa, Xavier Correig, Antonio Julià, and Sara Marsal. Focus: A Robust Workflow for One-Dimensional NMR Spectral Analysis. *Analytical Chemistry*, 86(2):1160–1169, 2013.

[5] Edward V Loftus. Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences. *Gastroenterology*, 126(6):1504–1517, 2004.

[6] Jacques Cosnes, Corinne Gower-Rousseau, Philippe Seksik, and Antoine Cortot. Epidemiology and natural history of inflammatory bowel diseases. *Gastroenterology*, 140 (6):1785–1794.e4, 2011.

[7] J. Satsangi, M. S. Silverberg, S. Vermeire, and J. F. Colombel. The montreal classification of inflammatory bowel disease: controversies, consensus, and implications. *Gut*, 55(6):749–753, 2006.

[8] T. M. Bayless, A. Z. Tokayer, J. M. Polito 2nd, S. A. Quaskey, E. D. Mellits, and M. L. Harris. Crohn's disease: Concordance for site and clinical type in affected family members–potential hereditary influences. *Gastroenterology*, 111(3):573–579, 1996.

[9] J. F. Colombel, B. Grandbastien, C. Gower-Rousseau, S. Plegat, J. P. Evrard, J. L. Dupas, J. P. Gendre, R. Modigliani, J. Belaiche, J. Hostein, J. P. Hugot, H. van Kruiningen, and A. Cortot. Clinical characteristics of Crohn's disease in 72 families. *Gastroenterology*, 111(3):604–607, 1996.

[10] Cecília Duraes, José C. Machado, Francisco Portela, Susana Rodrigues, Paula Lago, Marília Cravo, Paula Ministro, Margarida Marques, Isabelle Cremers, Joao Freitas, José Cotter, Lurdes Tavares, Leopoldo Matos, Isabel Medeiros, Rui Sousa, Jaime Ramos, Joao Deus, Paulo Caldeira, Cristina Chagas, Maria A. Duarte, Raquel Gonçalves, Rui Loureiro, Luísa Barros, Isabel Bastos, Eugénia Cancela, Mário C. Moraes, Maria J. Moreira, Ana I. Vieira, and Fernando Magro. Phenotype-genotype profiles in Crohn's disease predicted by genetic markers in autophagy-related genes (GOIA study ii). *Inflammatory Bowel Diseases*, 19(2):230–9, 2012.

[11] J. R. Fraser Cummings, R. M. Cooney, G. Clarke, J. Beckly, A. Geremia, S. Pathan, L. Hancock, C. Guo, L. R. Cardon, and D. P. Jewell. The genetics of NOD-like receptors in Crohn's disease. *Tissue Antigens*, 76(1):48–56, 2010.

[12] T. W. Eglinton, R. Roberts, J. Pearson, M. Barclay, T. R. Merriman, F. A. Frizelle, and R. B. Gearry. Clinical and genetic risk factors for perianal Crohn's disease in a population-based cohort. *Am J Gastroenterol*, 107(4):589–596, 2012.

[13] Camille Jung, Jean-Frédéric Colombel, Marc Lemann, Laurent Beaugerie, Matthieu Allez, Jacques Cosnes, Gwenola Vernier-Massouille, Jean-Marc Gornet, Jean-Pierre Gendre, Jean-Pierre Cezard, Frank M. Ruemmele, Dominique Turck, Françoise Merlin, Habib Zouali, Christian Libersa, Philippe Dieudé, Nadem Soufir, Gilles Thomas, and Jean-Pierre Hugot. Genotype/phenotype analyses for 53 Crohn's disease associated genetic polymorphisms. *PLoS ONE*, 7(12):e52223, 2012.

[14] James C Lee, Marion Espéli, Carl A Anderson, Michelle A Linterman, Joanna M Pocock, Naomi J Williams, Rebecca Roberts, Sebastien Viatte, Bo Fu, Norbert Peshu, Tran Tinh Hien, Nguyen Hoan Phu, Emma Wesley, Cathryn Edwards, Tariq Ahmad, John C Mansfield, Richard Gearry, Sarah Dunstan, Thomas N Williams, Anne Barton, Carola G Vinuesa, Anne Phillips, Craig Mowat, Hazel Drummond, Nick Kennedy, Charlie W Lees, Jack Satsangi, Kirstin Taylor, Natalie J Prescott, Christopher G Mathew, Peter Simpson, Alison Simmons, Mohammed Khan, William G Newman, Christopher Hawkey, Ailsa Hart, David C Wilson, Paul Henderson, Jeffrey C Barrett, Miles Parkes, Paul A. Lyons, and Kenneth G. C. Smith. Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. *Cell*, 155(1):57–69, 2013.

[15] Jeremy Adler, Sujal C. Rangwalla, Ben A. Dwamena, and Peter D. R. Higgins. The prognostic power of the NOD2 genotype for complicated Crohn's disease: A meta-analysis. *Am J Gastroenterol*, 106(4):699–712, 2011.

[16] Michael Economou, Thomas A. Trikalinos, Konstantinos T. Loizou, Epameinondas V. Tsianos, and John P. A. Ioannidis. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: A metaanalysis. *Am J Gastroenterol*, 99(12):2393–2404, 2004.

[17] Arnald Alonso, Eugeni Domènech, Antonio Julià, Julián Panés, Valle García-Sánchez, Pilar Nos Mateu, Ana Gutiérrez, Fernando Gomollón, Juan L Mendoza, Esther Garcia-Planella, et al. Identification of risk loci for Crohn's disease phenotypes using a genome-wide association study. *Gastroenterology*, 2014.

[18] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007. doi: 10.1038/nature05911.

[19] Joel N Hirschhorn and Zofia K Z Gajdos. Genome-wide association studies: results from the first few years and potential implications for clinical medicine. *Annual review of medicine*, 62:11–24, January 2011. ISSN 1545-326X. doi: 10.1146/annurev.med.091708.162036.

[20] Antonio Julià, Javier Ballina, Juan D Cañete, Alejandro Balsa, Jesus Tornero-Molina, Antonio Naranjo, Mercedes Alperi-López, Alba Erra, Dora Pascual-Salcedo, Pere Barceló, Jordi Camps, and Sara Marsal. Genome-wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility. *Arthritis & Rheumatism*, 58(8):2275–2286, 2008.

[21] Teri A Manolio, Lisa D Brooks, and Francis S Collins. A HapMap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation*, 118(5):1590–1605, 2008.

[22] Jiannis Ragoussis. Genotyping technologies for genetic research. *Annual review of genomics and human genetics*, 10:117–33, January 2009. ISSN 1545-293X. doi:

10.1146/annurev-genom-082908-150116.

[23] Ann-Christine Syvänen. Toward genome-wide snp genotyping. *Nature genetics*, 37: S5–S10, 2005.

[24] Donald F Conrad, Dalila Pinto, Richard Redon, and Lars Feuk. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–12, April 2010. ISSN 1476-4687. doi: 10.1038/nature08516.

[25] Richard Redon, Shumpei Ishikawa, Karen R Fitch, Lars Feuk, George H Perry, T Daniel Andrews, Heike Fiegler, Michael H Shapero, Andrew R Carson, Wenwei Chen, Eun Kyung Cho, Stephanie Dallaire, Jennifer L Freeman, Juan R Gonzalez, Monica Gratacos, Jing Huang, Dimitrios Kalaitzopoulos, Daisuke Komura, Jeffrey R MacDonald, Christian R Marshall, Rui Mei, Lyndal Montgomery, Kunihiro Nishimura, Kohji Okamura, Fan Shen, Martin J Somerville, Joelle Tchinda, Armand Valsesia, Cara Woodwark, Fengtang Yang, Junjun Zhang, Tatiana Zerjal, Jane Zhang, Lluis Armengol, Donald F Conrad, Xavier Estivill, Chris Tyler-Smith, Nigel P Carter, Hiroyuki Aburatani, Charles Lee, Keith W Jones, Stephen W Scherer, and Matthew E Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, November 2006. ISSN 0028-0836.

[26] Pawe Stankiewicz and James R Lupski. Structural variation in the human genome and its role in disease. *Annual review of medicine*, 61:437–55, January 2010. ISSN 1545-326X. doi: 10.1146/annurev-med-100708-204735.

[27] Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nat Rev Genet*, 7(2):85–97, 2006. doi: 10.1038/nrg1767.

[28] Peter H Sudmant, Jacob O Kitzman, Francesca Antonacci, Can Alkan, Maika Malig, Anya Tsalenko, Nick Sampas, Laurakay Bruhn, Jay Shendure, Project Genomes, and Evan E Eichler. Diversity of Human Copy Number Variation and Multicopy Genes. *Science*, 330(6004):641–646, 2010.

[29] Xavier Estivill and Lluís Armengol. Copy Number Variants and Common Disorders: Filling the Gaps and Exploring Complexity in Genome-Wide Association Studies. *PLoS Genet*, 3(10):e190, October 2007.

[30] Feng Zhang, Wenli Gu, Matthew E Hurles, and James R Lupski. Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, 10(1):451–481, 2009. doi: 10.1146/annurev.genom.9.081307.164217.

[31] Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464 (7289):713–720, 2010. doi: 10.1038/nature08979.

[32] E S Lander, L M Linton, and B Birren. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. ISSN 0028-0836.

[33] Francis S Collins, Ari Patrinos, Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, et al. New goals for the US human genome project: 1998-2003. *Science*, 282(5389):682–689, 1998.

[34] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[35] Leslie Pray. Discovery of DNA structure and function: Watson and crick. *Nature Education*, 1(1):100, 2008.

[36] Iakes Ezkurdia, David Juan, Jose Manuel Rodriguez, Adam Frankish, Mark Diekhans, Jennifer Harrow, Jesus Vazquez, Alfonso Valencia, and Michael L Tress. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding

genes. *Human molecular genetics*, 23(22):5866–5878, 2014.

[37] Hongzhu Qu and Xiangdong Fang. A brief review on the human encyclopedia of DNA elements (ENCODE) project. *Genomics, proteomics & bioinformatics*, 11(3):135–141, 2013.

[38] Susanne Hiller-Sturmhöfel. Gene structure and gene expression in higher organisms. *Alcohol Research*, 30(1):8, 2007.

[39] Roger P Alexander, Gang Fang, Joel Rozowsky, Michael Snyder, and Mark B Gerstein. Annotating non-coding regions of the genome. *Nat Rev Genet*, 11(8):559–571, 2010. ISSN 1471-0056.

[40] Annita Quintal Gomes, Sofia Nolasco, and Helena Soares. Non-Coding RNAs: Multi Tasking Molecules in the Cell. *International Journal of Molecular Sciences*, 14(8): 16010–16039, 2013.

[41] Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376, 2011. ISSN 1471-0056.

[42] Jacques S Beckmann, Xavier Estivill, and Stylianos E Antonarakis. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet*, 8(8):639–646, August 2007. ISSN 1471-0056.

[43] Jeffrey R MacDonald, Robert Ziman, Ryan KC Yuen, Lars Feuk, and Stephen W Scherer. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic acids research*, 42(D1):D986–D992, 2014.

[44] Joachim Weischenfeldt, Orsolya Symmons, Francois Spitz, and Jan O Korbel. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14(2):125–138, 2013.

[45] Alexander Martínez-Fundichely, Sònia Casillas, Raquel Egea, Miquel Ràmia, Antonio Barbadilla, Lorena Pantano, Marta Puig, and Mario Cáceres. InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic acids research*, page gkt1122, 2013.

[46] Jeffrey D Wall and Jonathan K Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8):587–597, 2003.

[47] Simon Myers, Leonardo Bottolo, Colin Freeman, Gil McVean, and Peter Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, 2005.

[48] Montgomery Slatkin. Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

[49] Chia-Ho Lin. Linkage disequilibrium measures. *Master of Science in Statistics. UCLA*, 2005.

[50] Daniel O Stram. Tag snp selection for association studies. *Genetic epidemiology*, 27 (4):365–374, 2004.

[51] The International HapMap Consortium. The International HapMap Project. *Nature*, 426(6968):789–96, December 2003. ISSN 1476-4687.

[52] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[53] West Africa. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, October 2010. ISSN 1476-4687. doi: 10.1038/nature09534.

[54] Eric Peacock and Phyllis Whiteley. Perlegen sciences, inc. *Pharmacogenomics*, 6(4): 439–442, July 2005. ISSN 1462-2416. doi: 10.1517/14622416.6.4.439.

[55] Giulia C Kennedy, Hajime Matsuzaki, Shoulian Dong, Wei-min Liu, Jing Huang, Guoy-

ing Liu, Xing Su, Manqiu Cao, Wenwei Chen, Jane Zhang, Weiwei Liu, Geoffrey Yang, Xiaojun Di, Thomas Ryder, Zhijun He, Urvashi Surti, Michael S Phillips, Michael T Boyce-Jacino, Stephen PA Fodor, and Keith W Jones. Large-scale genotyping of complex DNA. *Nat Biotech*, 21(10):1233–1237, October 2003. ISSN 1087-0156.

[56] Frank J Steemers, Weihua Chang, Grace Lee, David L Barker, Richard Shen, and Kevin L Gunderson. Whole-genome genotyping with the single-base extension assay. *Nat Meth*, 3(1):31–33, 2006. doi: 10.1038/nmeth842.

[57] William S Bush and Jason H Moore. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012.

[58] Kevin L Gunderson. Whole-genome genotyping on bead arrays. In *DNA Microarrays for Biomedical Research*, pages 197–213. Springer, 2009.

[59] Stephen F Kingsmore, Ingrid E Lindquist, Joann Mudge, Damian D Gessler, and William D Beavis. Genome-wide association studies: progress and potential for drug discovery and development. *Nature Reviews Drug Discovery*, 7(3):221–230, 2008.

[60] Beben Benyamin, Peter M Visscher, and Allan F McRae. Family-based genome-wide association studies. *Pharmacogenomics*, 10(2):181–190, 2009.

[61] Eric S Lander and Nicholas J Schork. Genetic dissection of complex traits. *Science*, 265(5181):2037–2048, 1994.

[62] Neil Risch and Jun Teng. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases i. DNA pooling. *Genome Research*, 8(12):1273–1288, 1998.

[63] Jun Teng and Neil Risch. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. ii. individual genotyping. *Genome Research*, 9(3):234–241, 1999.

[64] Danielle M Dick and Tatiana Foroud. Genetic strategies to detect genes involved in alcoholism and alcohol-related traits. *Alcohol Research*, 26(3):172, 2002.

[65] Robert Plomin, Claire MA Haworth, and Oliver SP Davis. Common disorders are quantitative traits. *Nature Reviews Genetics*, 10(12):872–878, 2009.

[66] David J Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[67] Geraldine M Clarke, Carl A Anderson, Fredrik H Pettersson, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133, 2011.

[68] Gustavo Barcelos Barra, Ludmila Alves, Silvia Conde, Patricia Godoy, Patricia Sales, Monalisa Ferreira, and Angelica Amorim. Association of the rs7903146 single nucleotide polymorphism at the Transcription Factor 7-like 2 (TCF7L2) locus with type 2 diabetes in Brazilian subjects. *Arquivos Brasileiros de Endocrinologia and Metabologia*, 56(8), 2012.

[69] Itsik Pe'er, Roman Yelensky, David Altshuler, and Mark J Daly. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic epidemiology*, 32(4):381–385, 2008.

[70] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

[71] Lin Hou and Hongyu Zhao. A review of post-GWAS prioritization approaches. *Frontiers in genetics*, 4, 2013.

[72] Illumina. Illumina GenCall Data Analysis Software. *Technology Spotlight*, 2005.

[73] Eleni Giannoulatou, Christopher Yau, Stefano Colella, Jiannis Ragoussis, and Christo-

pher C Holmes. GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. *Bioinformatics*, 24(19):2209–2214, 2008.

[74] Gengxin Li, Joel Gelernter, Henry R Kranzler, and Hongyu Zhao. M3: an improved SNP calling algorithm for Illumina BeadArray data. *Bioinformatics*, 28(3):358–365, 2012.

[75] Daniel A Peiffer, Jennie M Le, Frank J Steemers, Weihua Chang, Tony Jenniges, Francisco Garcia, Kirt Haden, Jiangzhen Li, Chad A Shaw, John Belmont, Sau Wai Cheung, Richard M Shen, David L Barker, and Kevin L Gunderson. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping . *Genome Research*, 16(9):1136–1148, September 2006.

[76] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F A Grant, Hakon Hakonarson, and Maja Bucan. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17(11):1665–1674, 2007.

[77] Stefano Colella, Christopher Yau, Jennifer M Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S Bassett, Anneke Seller, Christopher C Holmes, and Jiannis Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, 35(6):2013–2025, 2007.

[78] Dalila Pinto, Katayoon Darvishi, Xinghua Shi, Diana Rajan, Diane Rigler, Tom Fitzgerald, Anath C Lionel, Bhooma Thiruvahindrapuram, Jeffrey R Macdonald, Ryan Mills, Aparna Prasad, Kristin Noonan, Susan Gribble, Elena Prigmore, Patricia K Donahoe, Richard S Smith, Ji Hyeon Park, Matthew E Hurles, Nigel P Carter, Charles Lee, Stephen W Scherer, and Lars Feuk. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature biotechnology*, 29(6):512–20, January 2011. ISSN 1546-1696. doi: 10.1038/nbt.1852.

[79] Gary J. Patti, Oscar Yanes, and Gary Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*, 13(4):263–269, 2012.

[80] Xiaoquan Qi and Dabing Zhang. Plant metabolomics and metabolic biology. *Journal of Integrative Plant Biology*, 56(9):814–815, 2014.

[81] M. Oresic. Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction. *Nutrition, Metabolism and Cardiovascular Diseases*, 19(11):816–824, 2009.

[82] Helena Gibbons, Aoife O'Gorman, and Lorraine Brennan. Metabolomics as a tool in nutritional research. *Current Opinion in Lipidology*, 26(1), 2015.

[83] Christa Kühn. Metabolomics in animal breeding. In *Genetics Meets Metabolomics*, pages 107–123. Springer, 2012.

[84] DG Robertson and U Frevert. Metabolomics in drug discovery and development. *Clinical Pharmacology & Therapeutics*, 94(5):559–561, 2013.

[85] Douglas B Kell and Royston Goodacre. Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. *Drug discovery today*, 19(2):171–182, 2014.

[86] Rima Kaddurah-Daouk, Bruce S. Kristal, and Richard M. Weinshilboum. Metabolomics: A global biochemical approach to drug response and disease. *Annual Review of Pharmacology and Toxicology*, 48(1):653–683, 2008.

[87] Mamas Mamas, WarwickB Dunn, Ludwig Neyses, and Royston Goodacre. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Archives*

of Toxicology, 85(1):5–17, 2011.

[88] Urs A. Meyer, Ulrich M. Zanger, and Matthias Schwab. Omics and drug response. *Annual Review of Pharmacology and Toxicology*, 53(1):475–502, 2013.

[89] Emily G. Armitage and Coral Barbas. Metabolomics in cancer biomarker discovery: Current trends and future perspectives. *Journal of Pharmaceutical and Biomedical Analysis*, 87(0):1–11, 2014.

[90] Antonio Julià, Arnald Alonso, and Sara Marsal. Metabolomics in rheumatic diseases. *International Journal of Clinical Rheumatology*, 9(4):353–369, 2014.

[91] Tobias Fuhrer and Nicola Zamboni. High-throughput discovery metabolomics. *Current Opinion in Biotechnology*, 31(0):73–78, 2015.

[92] J. H. Bothwell and J. L. Griffin. An introduction to biological nuclear magnetic resonance spectroscopy. *Biol Rev Camb Philos Soc.*, 86(2):493–510, 2011. doi: 10.1111/j.1469-185X.2010.00157.x.

[93] Nicholas V. Reo. NMR-based metabolomics. *Drug and Chemical Toxicology*, 25(4): 375–382, 2002.

[94] Olaf Beckonert, Hector C Keun, Timothy MD Ebbels, Jacob Bundy, Elaine Holmes, John C Lindon, and Jeremy K Nicholson. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols*, 2(11):2692–2703, 2007.

[95] Jane L. Ward, John M. Baker, and Michael H. Beale. Recent applications of NMR spectroscopy in plant metabolomics. *FEBS Journal*, 274(5):1126–1131, 2007.

[96] Bernhard Blumich. Principles of nuclear magnetic resonance microscopy. *Magnetic Resonance in Chemistry*, 33(4):322–322, 1995.

[97] Wolfgang Dietrich, Christian H. Rudel, and Markus Neumann. Fast and precise automatic baseline correction of one- and two-dimensional NMR spectra. *Journal of Magnetic Resonance (1969)*, 91(1):1–11, 1991.

[98] Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. Xcms: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006.

[99] Yuanxin Xi and David Rocke. Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics*, 9(1):324, 2008.

[100] Zhi-Min Zhang, Shan Chen, and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146, 2010.

[101] David Chang, Cory D Banack, and Sirish L Shah. Robust baseline correction algorithm for signal dense NMR spectra. *Journal of Magnetic Resonance*, 187(2):288–292, 2007.

[102] Helen G. Gika, Georgios A. Theodoridis, Robert S. Plumb, and Ian D. Wilson. Current practice of liquid chromatography-mass spectrometry in metabolomics and metabonomics. *Journal of Pharmaceutical and Biomedical Analysis*, 87(0):12–25, 2014.

[103] Weihuan Niu, Elisa Knight, Qingyou Xia, and Brian D. McGarvey. Comparative evaluation of eight software programs for alignment of gas chromatography-mass spectrometry chromatograms in metabolomics experiments. *Journal of Chromatography A*, 1374(0):199–206, 2014.

[104] Atefeh Rafiei and Lekha Sleno. Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis. *Rapid Communications in Mass Spectrometry*, 29(1):119–127, 2015.

[105] David S. Wishart. Quantitative metabolomics using NMR. *TrAC Trends in Analytical Chemistry*, 27(3):228–237, 2008.

[106] Trung Vu and Kris Laukens. Getting your peaks in line: A review of alignment methods for NMR spectral data. *Metabolites*, 3(2):259–276, 2013.

[107] S. A. A. Sousa, Alviclér Magalhaes, and Márcia Miguel Castro Ferreira. Optimized bucketing for NMR spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems*, 122(0):93–102, 2013.

[108] Aalim M. Weljie, Jack Newton, Pascal Mercier, Erin Carlson, and Carolyn M. Slupsky. Targeted profiling: Quantitative analysis of 1H NMR metabolomics data. *Analytical Chemistry*, 78(13):4430–4442, 2006.

[109] Jie Hao, Manuel Liebeke, William Astle, Maria De Iorio, Jacob G. Bundy, and Timothy M. D. Ebbels. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using batman. *Nat. Protocols*, 9(6):1416–1427, 2014.

[110] R. Tautenhahn, C. Bottcher, and S. Neumann. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, 9:504, 2008.

[111] Ralf Tautenhahn, Gary J. Patti, Duane Rinehart, and Gary Siuzdak. Xcms online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, 84(11):5035–5039, 2012.

[112] Tomas Pluskal, Sandra Castillo, Alejandro Villar-Briones, and Matej Oresic. Mzmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11(1):395, 2010.

[113] Chao Yang, Zengyou He, and Weichuan Yu. Comparison of public peak detection algorithms for maldi mass spectrometry data analysis. *BMC Bioinformatics*, 10(1):4, 2009.

[114] P. Du, W. A. Kibbe, and S. M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22:2059–2065, 2006.

[115] C. Xiao, F. Hao, X. Qin, Y. Wang, and H. Tang. An optimized buffer system for NMR-based urinary metabonomics with effective ph control, chemical shift consistency and dilution minimization. *Analyst.*, 134(5):916–25, 2009. doi: 10.1039/b818802e.

[116] Lyle Burton, Gordana Ivosev, Stephen Tate, Gary Impey, Julie Wingate, and Ron Bonner. Instrumental and experimental effects in lc-ms-based metabolomics. *Journal of Chromatography B*, 871(2):227–235, 2008.

[117] Giorgio Tomasi, Frans van den Berg, and Claus Andersson. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, 18(5):231–241, 2004.

[118] Paul H. C. Eilers. Parametric time warping. *Analytical Chemistry*, 76(2):404–411, 2003.

[119] Jenny Forshed, Ina Schuppe-Koistinen, and Sven P. Jacobsson. Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta*, 487(2):189–199, 2003.

[120] Geun-Cheol Lee and David L. Woodruff. Beam search for peak alignment of NMR signals. *Analytica Chimica Acta*, 513(2):413–416, 2004.

[121] David Clifford, Glenn Stone, Ivan Montoliu, Serge Rezzi, François-Pierre Martin, Philippe Guy, Stephen Bruce, and Sunil Kochhar. Alignment using variable penalty dynamic time warping. *Analytical Chemistry*, 81(3):1000–1007, 2009.

[122] F. Savorani, G. Tomasi, and S. B. Engelsen. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2):190–202, 2010.

[123] Jason W. H. Wong, Caterina Durante, and Hugh M. Cartwright. Application of fast fourier transform cross-correlation for the alignment of large chromatographic and

spectral datasets. *Analytical Chemistry*, 77(17):5655–5661, 2005.

[124] Kirill A. Veselkov, John C. Lindon, Timothy M. D. Ebbels, Derek Crockford, Vladimir V. Volynkin, Elaine Holmes, David B. Davies, and Jeremy K. Nicholson. Recursive segment-wise peak alignment of biological 1H NMR spectra for improved metabolic biomarker recovery. *Analytical Chemistry*, 81(1):56–66, 2008.

[125] Guro F. Giskeodegard, Tom G. Bloemberg, Geert Postma, Beathe Sitter, May-Britt Tessem, Ingrid S. Gribbestad, Tone F. Bathen, and Lutgarde M. C. Buydens. Alignment of high resolution magic angle spinning magnetic resonance spectra using warping methods. *Analytica Chimica Acta*, 683(1):1–11, 2010.

[126] Wei Jiang, Zhi-Min Zhang, YongHuan Yun, De-Jian Zhan, Yi-Bao Zheng, Yi-Zeng Liang, ZhenYu Yang, and Ling Yu. Comparisons of five algorithms for chromatogram alignment. *Chromatographia*, 76(17-18):1067–1078, 2013.

[127] A. M. van Nederkassel, M. Daszykowski, P. H. C. Eilers, and Y. Vander Heyden. A comparison of three algorithms for chromatograms alignment. *Journal of Chromatography A*, 1118(2):199–210, 2006.

[128] Ralf Tautenhahn, Kevin Cho, Winnie Uritboonthai, Zhengjiang Zhu, Gary J. Patti, and Gary Siuzdak. An accelerated workflow for untargeted metabolomics using the metlin database. *Nat Biotech*, 30(9):826–828, 2012.

[129] Andrew Craig, Olivier Cloarec, Elaine Holmes, Jeremy K. Nicholson, and John C. Lindon. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7):2262–2267, 2006.

[130] StefanieM Kohl, MatthiasS Klein, Jochen Hochrein, PeterJ Oefner, Rainer Spang, and Wolfram Gronwald. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, 8(1):146–160, 2012.

[131] LoneG Rasmussen, Francesco Savorani, ThomasM Larsen, LarsO Dragsted, Arne Astrup, and SorenB Engelsen. Standardization of factors that influence human urine metabolomics. *Metabolomics*, 7(1):71–83, 2011.

[132] Arnald Alonso, Antonio Julià, Antoni Beltran, Maria Vinaixa, Marta Díaz, Lourdes Ibanez, Xavier Correig, and Sara Marsal. AStream: An R Package for Annotating LC/MS Metabolomic Data. *Bioinformatics*, 27(9):1339–1340, 2011.

[133] Carsten Kuhl, Ralf Tautenhahn, Christoph Bottcher, Tony R. Larson, and Steffen Neumann. Camera: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84(1):283–289, 2011.

[134] Dan Tulpan, Serge Leger, Luc Belliveau, Adrian Culf, and Miroslava Cuperlovic-Culf. Metabohunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC Bioinformatics*, 12(1):400, 2011.

[135] Pascal Mercier, MichaelJ Lewis, David Chang, David Baker, and DavidS Wishart. Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra. *Journal of Biomolecular NMR*, 49(3-4):307–323, 2011.

[136] Daniel Jacob, Catherine Deborde, and Annick Moing. An efficient spectra processing method for metabolite identification from 1H-NMR metabolomics data. *Analytical and Bioanalytical Chemistry*, 405(15):5049–5061, 2013.

[137] James J. Ellinger, Roger A. Chylla, Eldon L. Ulrich, and John L. Markley. Databases and software for NMR-based metabolomics. *Current Metabolomics*, 1(1): 10.2174/2213235X11301010028, 2013.

[138] Atsushi Fukushima and Miyako Kusano. Recent progress in the development of metabolome databases for plant systems biology. *Frontiers in Plant Science*, 4:73,

2013.

[139] David S. Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, Souhaila Bouatra, Igor Sinelnikov, David Arndt, Jianguo Xia, Philip Liu, Faizath Yallou, Trent Bjorndahl, Rolando Perez-Pineiro, Roman Eisner, Felicity Allen, Vanessa Neveu, Russ Greiner, and Augustin Scalbert. HMDB 3.0: The human metabolome database in 2013. *Nucleic Acids Research*, 41(D1):D801–D807, 2013.

[140] Nikolaos Psychogios, David D. Hau, Jun Peng, An Chi Guo, Rupasri Mandal, Souhaila Bouatra, Igor Sinelnikov, Ramanarayan Krishnamurthy, Roman Eisner, Bijaya Gautam, Nelson Young, Jianguo Xia, Craig Knox, Edison Dong, Paul Huang, Zsuzsanna Hollander, Theresa L. Pedersen, Steven R. Smith, Fiona Bamforth, Russ Greiner, Bruce McManus, John W. Newman, Theodore Goodfriend, and David S. Wishart. The human serum metabolome. *PLoS ONE*, 6(2):e16957, 2011.

[141] Souhaila Bouatra, Farid Aziat, Rupasri Mandal, An Chi Guo, Michael R. Wilson, Craig Knox, Trent C. Bjorndahl, Ramanarayan Krishnamurthy, Fozia Saleem, Philip Liu, Zerihun T. Dame, Jenna Poelzer, Jessica Huynh, Faizath S. Yallou, Nick Psychogios, Edison Dong, Ralf Bogumil, Cornelia Roehring, and David S. Wishart. The human urine metabolome. *PLoS ONE*, 8(9):e73076, 2013.

[142] Catarina D Campbell, Nick Sampas, Anya Tsalenko, Peter H Sudmant, Jeffrey M Kidd, Maika Malig, Tiffany H Vu, Laura Vives, Peter Tsang, Laurakay Bruhn, and Evan E Eichler. Population-Genetic Properties of Differentiated Human Copy-Number Polymorphisms. *The American Journal of Human Genetics*, 88(3):317–332, 2011. doi: 10.1016/j.ajhg.2011.02.004.

[143] Steven A McCarroll, Finny G Kuruvilla, Joshua M Korn, Simon Cawley, and David Altshuler. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*, 40(10):1166–1174, October 2008. ISSN 1061-4036.

[144] Josephine Elia, Joseph T Glessner, Kai Wang, Nagahide Takahashi, Corina J Shtir, Dexter Hadley, Patrick M A Sleiman, Haitao Zhang, Cecilia E Kim, Reid Robison, Gholson J Lyon, James H Flory, Jonathan P Bradfield, Marcin Imielinski, Cuiping Hou, Edward C Frackelton, Rosetta M Chiavacci, Takeshi Sakurai, Cara Rabin, Frank A Middleton, Kelly A Thomas, Maria Garris, Frank Mentch, Christine M Freitag, Hans-Christoph Steinhausen, Alexandre A Todorov, Andreas Reif, Aribert Rothenberger, Barbara Franke, Eric O Mick, Herbert Roeyers, Jan Buitelaar, Klaus-Peter Lesch, Tobias Banaschewski, Richard P Ebstein, Fernando Mulas, Robert D Oades, Joseph Sergeant, Edmund Sonuga-Barke, Tobias J Renner, Marcel Romanos, Jasmin Romanos, Andreas Warnke, Susanne Walitza, Jobst Meyer, Haukur Palmason, Christiane Seitz, Sandra K Loo, Susan L Smalley, Joseph Biederman, Lindsey Kent, Philip Asherson, Richard J L Anney, J William Gaynor, Philip Shaw, Marcella Devoto, Peter S White, Struan F A Grant, Joseph D Buxbaum, Judith L Rapoport, Nigel M Williams, Stanley F Nelson, Stephen V Faraone, and Hakon Hakonarson. Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder. *Nat Genet*, 44(1):78–84, January 2012. ISSN 1061-4036.

[145] Joseph T Glessner, Kai Wang, Guiqing Cai, Olena Korvatska, Cecilia E Kim, Shawn Wood, Haitao Zhang, Annette Estes, Camille W Brune, Jonathan P Bradfield, Marcin Imielinski, Edward C Frackelton, Jennifer Reichert, Emily L Crawford, Jeffrey Munson, Patrick M A Sleiman, Rosetta Chiavacci, Kiran Annaiah, Kelly Thomas, Cuiping Hou, Wendy Glaberson, James Flory, Frederick Otieno, Maria Garris, Latha Soorya,

Lambertus Klei, Joseph Piven, Kacie J Meyer, Evdokia Anagnostou, Takeshi Sakurai, Rachel M Game, Danielle S Rudd, Danielle Zurawiecki, Christopher J McDougle, Lea K Davis, Judith Miller, David J Posey, Shana Michaels, Alexander Kolevzon, Jeremy M Silverman, Raphael Bernier, Susan E Levy, Robert T Schultz, Geraldine Dawson, Thomas Owley, William M McMahon, Thomas H Wassink, John A Sweeney, John I Nurnberger, Hilary Coon, James S Sutcliffe, Nancy J Minshew, Struan F A Grant, Maja Bucan, Edwin H Cook, Joseph D Buxbaum, Bernie Devlin, Gerard D Schellenberg, and Hakon Hakonarson. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, 459(7246):569–573, May 2009. ISSN 0028-0836.

[146] Heather C Mefford, Hiltrud Muhle, Philipp Ostertag, Sarah von Spiczak, Karen Buysse, Carl Baker, Andre Franke, Alain Malafosse, Pierre Genton, Pierre Thomas, Christina A Gurnett, Stefan Schreiber, Alexander G Bassuk, Michel Guipponi, Ulrich Stephani, Ingo Helbig, and Evan E Eichler. Genome-Wide Copy Number Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies. *PLoS Genet*, 6(5):e1000962, May 2010.

[147] Nathan Day, Andrew Hemmaplardh, Robert E Thurman, John A Stamatoyannopoulos, and William S Noble. Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11):1424–1426, 2007.

[148] Arnald Alonso, Antonio Julia, Raul Tortosa, Cristina Canaleta, Juan Canete, Javier Ballina, Alejandro Balsa, Jesus Tornero, and Sara Marsal. CNstream: A method for the identification and genotyping of copy number polymorphisms using Illumina microarrays. *BMC Bioinformatics*, 11(1):264, 2010.

[149] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[150] Joanna Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh. McKusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Research*, 37(suppl 1):D793–D796, 2009.

[151] Eitan Halper-Stromberg, Laurence Frelin, Ingo Ruczinski, Robert Scharpf, Chunfa Jie, Benilton Carvalho, Haiping Hao, Kurt Hetrick, Anne Jedlicka, Amanda Dziedzic, Kim Doheny, Alan F Scott, Steve Baylin, Jonathan Pevsner, Forrest Spencer, and Rafael A Irizarry. Performance assessment of copy number microarray platforms using a spike-in experiment. *Bioinformatics*, 27(8):1052–1060, 2011.

[152] A P Dempster, N M Laird, and D B Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[153] Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Rolf N. Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov, and Alexandra Soboleva. NCBI GEO: Archive for Functional Genomics Data Sets - 10 years on. *Nucleic Acids Research*, 39(suppl 1):D1005–D1010, 2011.

[154] Rafael de Cid, Eva Riveira-Munoz, Patrick L J M Zeeuwen, Jason Robarge, Wilson Liao, Emma N Dannhauser, Emiliano Giardina, Philip E Stuart, Rajan Nair, Cynthia Helms, Georgia Escaramis, Ester Ballana, Gemma Martin-Ezquerra, Martin den Heijer, Marijke Kamsteeg, Irma Joosten, Evan E Eichler, Conxi Lazaro, Ramon M Pujol, Lluis Armengol, Goncalo Abecasis, James T Elder, Giuseppe Novelli, John A L Ar-

mour, Pui-Yan Kwok, Anne Bowcock, Joost Schalkwijk, and Xavier Estivill. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet*, 41(2):211–215, 2009. doi: 10.1038/ng.313.

[155] Steven A McCarroll, Alan Huett, Petric Kuballa, Shannon D Chilewski, Aimee Landry, Philippe Goyette, Michael C Zody, Jennifer L Hall, Steven R Brant, Judy H Cho, Richard H Duerr, Mark S Silverberg, Kent D Taylor, John D Rioux, David Altshuler, Mark J Daly, and Ramnik J Xavier. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet*, 40(9):1107–1112, 2008. doi: 10.1038/ng.215.

[156] Pauline A Fujita, Brooke Rhead, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Melissa S Cline, Mary Goldman, Galt P Barber, Hiram Clawson, Antonio Coelho, Mark Diekhans, Timothy R Dreszer, Belinda M Giardine, Rachel A Harte, Jennifer Hillman-Jackson, Fan Hsu, Vanessa Kirkup, Robert M Kuhn, Katrina Learned, Chin H Li, Laurence R Meyer, Andy Pohl, Brian J Raney, Kate R Rosenbloom, Kayla E Smith, David Haussler, and W James Kent. The UCSC Genome Browser database: update 2011. *Nucleic Acids Research*, 2010.

[157] Matthew E Ritchie, Ruijie Liu, Benilton S Carvalho, and Rafael a Irizarry. Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. *BMC bioinformatics*, 12(1):68, January 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-68.

[158] Yi Yu, Tushar R Bhangale, Jesen Fagerness, Stephan Ripke, Gudmar Thorleifsson, Perciliz L Tan, Eric H Souied, Andrea J Richardson, Joanna E Merriam, Gabriëlle H S Buitendijk, Robyn Reynolds, Soumya Raychaudhuri, Kimberly A Chin, Lucia Sobrin, Evangelos Evangelou, Phil H Lee, Aaron Y Lee, Nicolas Leveziel, Donald J Zack, Betsy Campochiaro, Peter Campochiaro, R Theodore Smith, Gaetano R Barile, Robyn H Guymer, Ruth Hogg, Usha Chakravarthy, Luba D Robman, Omar Gustafsson, Haraldur Sigurdsson, Ward Ortmann, Timothy W Behrens, Kari Stefansson, André G Uitterlinden, Cornelia M van Duijn, Johannes R Vingerling, Caroline C W Klaver, Rando Allikmets, Milam A Brantley, Paul N Baird, Nicholas Katsanis, Unnur Thorsteinsdottir, John P A Ioannidis, Mark J Daly, Robert R Graham, and Johanna M Seddon. Common variants near FRK/COL10A1 and VEGFA are associated with advanced age-related macular degeneration. *Human Molecular Genetics*, 20(18):3699–3709, 2011.

[159] Elizabeth K Speliotes, Cristen J Willer, Gonneke Willemsen, Daniel R Witte, Jacqueline C Witteman, Jianfeng Xu, Qunyuan Zhang, Lina Zgaga, Andreas Ziegler, Paavo Zitting, John P Beilby, I Sadaf Farooqi, Johannes Hebebrand, Heikki V Huikuri, Alan L James, Mika Kahonen, Douglas F Levinson, Fabio Macciardi, Markku S Nieminen, Claes Ohlsson, Lyle J Palmer, Paul M Ridker, Michael Stumvoll, Jacques S Beckmann, Heiner Boeing, Eric Boerwinkle, Dorret I Boomsma, Mark J Caulfield, Stephen J Chanock, Francis S Collins, L Adrienne Cupples, George Davey Smith, Jeanette Erdmann, Philippe Froguel, Henrik Gronberg, Ulf Gyllensten, Per Hall, Torben Hansen, Tamara B Harris, Andrew T Hattersley, Richard B Hayes, Joachim Heinrich, Frank B Hu, Kristian Hveem, Thomas Illig, Marjo-Riitta Jarvelin, Jaakko Kaprio, Fredrik Karpe, Kay-Tee Khaw, Lambertus A Kiemeney, Heiko Krude, Markku Laakso, Debbie A Lawlor, Andres Metspalu, Patricia B Munroe, Willem H Ouwehand, Oluf Pedersen, Brenda W Penninx, Annette Peters, Peter P Pramstaller, Thomas Quertermous, Thomas Reinehr, Aila Rissanen, Igor Rudan, Nilesh J Samani, Peter E H Schwarz, Alan R Shuldiner, Timothy D Spector, Jaakko Tuomilehto, Manuela Uda, Andre Uitterlinden, Timo T Valle, Martin Wabitsch, Gerard Waeber, Nicholas J Wareham, Hugh

Watkins, James F Wilson, Alan F Wright, M Carola Zillikens, Nilanjan Chatterjee, Steven A McCarroll, Shaun Purcell, Eric E Schadt, Peter M Visscher, Themistocles L Assimes, Ingrid B Borecki, Panos Deloukas, Caroline S Fox, Leif C Groop, Talin Haritunians, David J Hunter, Robert C Kaplan, Karen L Mohlke, Jeffrey R O'Connell, Leena Peltonen, David Schlessinger, David P Strachan, Cornelia M van Duijn, H Erich Wichmann, Timothy M Frayling, Unnur Thorsteinsdottir, Goncalo R Abecasis, Ines Barroso, Michael Boehnke, Kari Stefansson, Kari E North, Mark I McCarthy, Joel N Hirschhorn, Erik Ingelsson, and Ruth J F Loos. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11): 937–948, 2010. doi: 10.1038/ng.686.

[160] Chikashi Terao, Ryo Yamada, Koichiro Ohmura, Meiko Takahashi, Takahisa Kawaguchi, Yuta Kochi, Yukinori Okada, Yusuke Nakamura, Kazuhiko Yamamoto, Inga Melchers, Mark Lathrop, Tsuneyo Mimori, and Fumihiko Matsuda. The human AIRE gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. *Human Molecular Genetics*, 20(13): 2680–2685, 2011.

[161] Nicole Soranzo, Serena Sanna, Eleanor Wheeler, Christian Gieger, Dörte Radke, Josée Dupuis, Nabila Bouatia-Naji, Claudia Langenberg, Inga Prokopenko, Elliot Stolerman, Manjinder S Sandhu, Matthew M Heeney, Joseph M Devaney, Muredach P Reilly, and Sally L Ricketts. Common Variants at 10 Genomic Loci Influence Hemoglobin A1C Levels via Glycemic and Nonglycemic Pathways. *Diabetes*, 59(12): 3229–3239, 2010.

[162] Yoon Shin Cho, Chien-Hsiun Chen, Cheng Hu, Jirong Long, Rick Twee Hee Ong, Xueling Sim, Fumihiko Takeuchi, Ying Wu, Min Jin Go, Toshimasa Yamauchi, Yi-Cheng Chang, Soo Heon Kwak, Ronald C W Ma, Ken Yamamoto, Linda S Adair, Tin Aung, Qiuyin Cai, Li-Ching Chang, Yuan-Tsong Chen, Yutang Gao, Frank B Hu, Hyung-Lae Kim, Sangsoo Kim, Young Jin Kim, Jeannette Jen-Mai Lee, Nanette R Lee, Yun Li, Jian Jun Liu, Wei Lu, Jiro Nakamura, Eitaro Nakashima, Daniel Peng-Keat Ng, Wan Ting Tay, Fuu-Jen Tsai, Tien Yin Wong, Mitsuhiro Yokota, Wei Zheng, Rong Zhang, Congrong Wang, Wing Yee So, Keizo Ohnaka, Hiroshi Ikegami, Kazuo Hara, Young Min Cho, Nam H Cho, Tien-Jyun Chang, Yuqian Bao, Asa K Hedman, Andrew P Morris, Mark I McCarthy, Ryoichi Takayanagi, Kyong Soo Park, Weiping Jia, Lee-Ming Chuang, Juliana C N Chan, Shiro Maeda, Takashi Kadowaki, Jong-Young Lee, Jer-Yuarn Wu, Yik Ying Teo, E Shyong Tai, Xiao Ou Shu, Karen L Mohlke, Norihiro Kato, Bok-Ghee Han, and Mark Seielstad. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet*, 44(1): 67–72, 2012. doi: 10.1038/ng.1019.

[163] Ying Jin, Stanca A Birlea, Pamela R Fain, Katherine Gowan, Sheri L Riccardi, Paulene J Holland, Christina M Mailloux, Alexandra J D Sufit, Saunie M Hutton, Anita Amadi-Myers, Dorothy C Bennett, Margaret R Wallace, Wayne T McCormack, E Helen Kemp, David J Gawkrodger, Anthony P Weetman, Mauro Picardo, Giovanni Leone, Alain Taïeb, Thomas Jouary, Khaled Ezzedine, Nanny van Geel, Jo Lambert, Andreas Overbeck, and Richard A Spritz. Variant of TYR and Autoimmunity Susceptibility Loci in Generalized Vitiligo. *New England Journal of Medicine*, 362(18):1686–1697, 2010. doi: 10.1056/NEJMoa0908547.

[164] Hamdi Mbarek, Hidenori Ochi, Yuji Urabe, Vinod Kumar, Michiaki Kubo, Naoya Hosono, Atsushi Takahashi, Yoichiro Kamatani, Daiki Miki, Hiromi Abe, Tatsuhiko Tsunoda, Naoyuki Kamatani, Kazuaki Chayama, Yusuke Nakamura, and Koichi Mat-

suda. A genome-wide association study of chronic hepatitis B identified novel risk locus in a Japanese population. *Human Molecular Genetics*, 20(19):3884–3892, 2011.

[165] John C Chambers, Weihua Zhang, Joban Sehmi, Xinzhong Li, Mark N Wass, Pim Van der Harst, Hilma Holm, Serena Sanna, Maryam Kavousi, Sebastian E Baumeister, Lachlan J Coin, Guohong Deng, Christian Gieger, Nancy L Heard-Costa, Jouke-Jan Hottenga, Brigitte Kuhnel, Vinod Kumar, Vasiliki Lagou, Liming Liang, Jian'an Luan, Pedro Marques Vidal, Irene Mateo Leach, Paul F O'Reilly, John F Peden, Nilufer Rahmioglu, Pasi Soininen, Elizabeth K Speliotes, Xin Yuan, Gudmar Thorleifsson, Behrooz Z Alizadeh, Larry D Atwood, Ingrid B Borecki, Morris J Brown, Pimphen Charoen, Francesco Cucca, Debashish Das, Eco J C de Geus, Anna L Dixon, Angela Doring, Georg Ehret, Gudmundur I Eyjolfsson, Martin Farrall, Nita G Forouhi, Nele Friedrich, Wolfram Goessling, Daniel F Gudbjartsson, Tamara B Harris, Anna-Liisa Hartikainen, Simon Heath, Gideon M Hirschfield, Albert Hofman, Georg Homuth, Elina Hypponen, Harry L A Janssen, Toby Johnson, Antti J Kangas, Ido P Kema, Jens P Kuhn, Sandra Lai, Mark Lathrop, Markus M Lerch, Yun Li, T Jake Liang, Jing-Ping Lin, Ruth J F Loos, Nicholas G Martin, Miriam F Moffatt, Grant W Montgomery, Patricia B Munroe, Kiran Musunuru, Yusuke Nakamura, Christopher J O'Donnell, Isleifur Olafsson, Brenda W Penninx, Anneli Pouta, Bram P Prins, Inga Prokopenko, Ralf Puls, Aimo Ruokonen, Markku J Savolainen, David Schlessinger, Jeoffrey N L Schouten, Udo Seedorf, Srijita Sen-Chowdhry, Katherine A Siminovitch, Johannes H Smit, Timothy D Spector, Wenting Tan, Tanya M Teslovich, Taru Tukiainen, Andre G Uitterlinden, Melanie M Van der Klauw, Ramachandran S Vasan, Chris Wallace, Henri Wallaschofski, H Erich Wichmann, Gonneke Willemsen, Peter Wurtz, Chun Xu, Laura M Yerges-Armstrong, Goncalo R Abecasis, Kourosh R Ahmadi, Dorret I Boomsma, Mark Caulfield, William O Cookson, Cornelia M van Duijn, Philippe Froguel, Koichi Matsuda, Mark I McCarthy, Christa Meisinger, Vincent Mooser, Kirsi H Pietilainen, Gunter Schumann, Harold Snieder, Michael J E Sternberg, Ronald P Stolk, Howard C Thomas, Unnur Thorsteinsdottir, Manuela Uda, Gerard Waeber, Nicholas J Wareham, Dawn M Waterworth, Hugh Watkins, John B Whitfield, Jacqueline C M Witteman, Bruce H R Wolffenbuttel, Caroline S Fox, Mika Ala-Korpela, Kari Stefansson, Peter Vollenweider, Henry Volzke, Eric E Schadt, James Scott, Marjo-Riitta Jarvelin, Paul Elliott, and Jaspal S Kooner. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet*, 43(11):1131–1138, 2011. doi: 10.1038/ng.970.

[166] Santhi K Ganesh, Neil A Zakai, Frank J A van Rooij, Nicole Soranzo, Albert V Smith, Michael A Nalls, Ming-Huei Chen, Anna Kottgen, Nicole L Glazer, Abbas Dehghan, Brigitte Kuhnel, Thor Aspelund, Qiong Yang, Toshiko Tanaka, Andrew Jaffe, Joshua C M Bis, Germaine C Verwoert, Alexander Teumer, Caroline S Fox, Jack M Guralnik, Georg B Ehret, Kenneth Rice, Janine F Felix, Augusto Rendon, Gudny Eiriksdottir, Daniel Levy, Kushang V Patel, Eric Boerwinkle, Jerome I Rotter, Albert Hofman, Jennifer G Sambrook, Dena G Hernandez, Gang Zheng, Stefania Bandinelli, Andrew B Singleton, Josef Coresh, Thomas Lumley, Andre G Uitterlinden, Janine M VanGils, Lenore J Launer, L Adrienne Cupples, Ben A Oostra, Jaap-Jan Zwaginga, Willem H Ouwehand, Swee-Lay Thein, Christa Meisinger, Panos Deloukas, Matthias Nauck, Tim D Spector, Christian Gieger, Vilmundur Gudnason, Cornelia M van Duijn, Bruce M Psaty, Luigi Ferrucci, Aravinda Chakravarti, Andreas Greinacher, Christopher J O'Donnell, Jacqueline C M Witteman, Susan Furth, Mary Cushman, Tamara B Harris, and Jing-Ping Lin. Multiple loci influence erythrocyte phenotypes in the

CHARGE Consortium. *Nat Genet*, 41(11):1191–1198, 2009. doi: 10.1038/ng.466.

[167] Yuan Jiang and Heping Zhang. Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder. *Genetic Epidemiology*, 35(2):125–132, 2011.

[168] Anne E Hughes, Nick Orr, Hossein Esfandiary, Martha Diaz-Torres, Timothy Goodship, and Usha Chakravarthy. A common CFH haplotype, with deletion of CFHR1 and CFHR3, is associated with lower risk of age-related macular degeneration. *Nat Genet*, 38(10):1173–1177, 2006. doi: 10.1038/ng1890.

[169] N R Rodrigues, N Owen, K Talbot, J Ignatius, V Dubowitz, and K E Davies. Deletions in the survival motor neuron gene on 5q13 in autosomal recessive spinal muscular atrophy. *Human Molecular Genetics*, 4(4):631–634, 1995.

[170] O R Adetunji, J C Blair, M Javadpour, A Alfirevic, M Pirmohamed, and I A MacFarlane. Deletion of exon 3 in the growth hormone receptor gene in adults with growth hormone deficiency: comparison of symptomatic and asymptomatic patients. *Clinical Endocrinology*, 72(3):422–423, 2010.

[171] C Bergmann, F Küpper, C P Schmitt, U Vester, T J Neuhaus, J Senderek, and K Zerres. Multi-exon deletions of the PKHD1 gene cause autosomal recessive polycystic kidney disease (ARPKD). *Journal of Medical Genetics*, 42(10):e63–e63, 2005.

[172] Illumina. Genome-Wide DNA Analysis BeadChips. *Illumnia Data Sheet*, 2010.

[173] Craddock N Moskvina V Holmans P, Owen MJ, O'Donovan MC. Effects of Differential Genotyping Error Rate on the Type I Error Probability of Case-Control Studies. *Human Heredity*, 61(1):55–64, 2006.

[174] Abecasis G R McCarthy MI Cardon LR, Goldstein DB, Little J. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9:356–369, 2008.

[175] Michael Nothnagel, David Ellinghaus, Stefan Schreiber, Michael Krawczak, and Andre Franke. A comprehensive evaluation of SNP genotype imputation. *Human Genetics*, 125(2):163–171, 2009. doi: 10.1007/s00439-008-0606-5.

[176] Lucy Huang, Yun Li, Andrew B Singleton, John A Hardy, Gonçalo Abecasis, Noah A Rosenberg, and Paul Scheet. Genotype-Imputation Accuracy across Worldwide Human Populations. *The American Journal of Human Genetics*, 84(2):235–250, 2009. doi: 10.1016/j.ajhg.2009.01.013.

[177] Eric R Gamazon, Dan L Nicolae, and Nancy J Cox. A Study of CNVs As Trait-Associated Polymorphisms and As Expression Quantitative Trait Loci. *PLoS Genet*, 7(2):e1001292, 2011. doi: 10.1371/journal.pgen.1001292.

[178] Santhosh Girirajan, Catarina D Campbell, and Evan E Eichler. Human Copy Number Variation and Complex Genetic Disease. *Annual Review of Genetics*, 45(1):203–226, 2011. doi: 10.1146/annurev-genet-102209-163544.

[179] Anna C Need, Dongliang Ge, Michael E Weale, Jessica Maia, Sheng Feng, Erin L Heinzen, Kevin V Shianna, Woohyun Yoon, Dalia Kasperaviciut, Massimo Gennarelli, Warren J Strittmatter, Cristian Bonvicini, Giuseppe Rossi, Karu Jayathilake, Philip A Cola, Joseph P McEvoy, Richard S E Keefe, Elizabeth M C Fisher, Pamela L St. Jean, Ina Giegling, Annette M Hartmann, Hans-Jürgen Möller, Andreas Ruppert, Gillian Fraser, Caroline Crombie, Lefkos T Middleton, David St. Clair, Allen D Roses, Pierandrea Muglia, Clyde Francks, Dan Rujescu, Herbert Y Meltzer, and David B Goldstein. A Genome-Wide Investigation of SNPs and CNVs in Schizophrenia. *PLoS Genet*, 5(2): e1000373, 2009. doi: 10.1371/journal.pgen.1000373.

[180] Nathan Pankratz, Alexandra Dumitriu, Kurt N Hetrick, Mei Sun, Jeanne C Latourelle,

Jemma B Wilk, Cheryl Halter, Kimberly F Doheny, James F Gusella, William C Nichols, Richard H Myers, Tatiana Foroud, Anita L DeStefano, PsgProgeni The, Coordinators GenePd Investigators, and Laboratories Molecular Genetic. Copy Number Variation in Familial Parkinson Disease. *PLoS ONE*, 6(8):e20988, 2011. doi: 10.1371/journal.pone.0020988.

[181] Clara Sze-Man Tang, Guo Cheng, Man-Ting So, Benjamin Hon-Kei Yip, Xiao-Ping Miao, Emily Hoi-Man Wong, Elly Sau-Wai Ngan, Vincent Chi-Hang Lui, You-Qiang Song, Danny Chan, Kenneth Cheung, Zhen-Wei Yuan, Liu Lei, Patrick Ho-Yu Chung, Xue-Lai Liu, Kenneth Kak-Yuen Wong, Christian R Marshall, Steve Scherer, Stacey S Cherny, Pak-Chung Sham, Paul Kwong-Hang Tam, and Maria-Mercè Garcia-Barceló. Genome-Wide Copy Number Analysis Uncovers a New HSCR Gene: <italic>NRG3</italic>. *PLoS Genet*, 8(5):e1002687, 2012. doi: 10.1371/journal.pgen.1002687.

[182] Jian Zhao, Hui Wu, Melanie Khosravi, Huijuan Cui, Xiaoxia Qian, Jennifer A Kelly, Kenneth M Kaufman, Carl D Langefeld, Adrienne H Williams, Mary E Comeau, Julie T Ziegler, Miranda C Marion, Adam Adler, Stuart B Glenn, Marta E Alarcón-Riquelme, Bernardo A Pons-Estel, John B Harley, Sang-Cheol Bae, So-Young Bang, Soo-Kyung Cho, Chaim O Jacob, Timothy J Vyse, Timothy B Niewold, Patrick M Gaffney, Kathy L Moser, Robert P Kimberly, Jeffrey C Edberg, Elizabeth E Brown, Graciela S Alarcon, Michelle A Petri, Rosalind Ramsey-Goldman, Luis M Vilá, John D Reveille, Judith A James, Gary S Gilkeson, Diane L Kamen, Barry I Freedman, Juan-Manuel Anaya, Joan T Merrill, Lindsey A Criswell, R Hal Scofield, Anne M Stevens, Joel M Guthridge, Deh-Ming Chang, Yeong Wook Song, Ji Ah Park, Eun Young Lee, Susan A Boackle, Jennifer M Grossman, Bevra H Hahn, Timothy H J Goodship, Rita M Cantor, Chack-Yung Yu, Nan Shen, Betty P Tsao, Biolupus Network, and Genles Network. Association of Genetic Variants in Complement Factor H and Factor H-Related Genes with Systemic Lupus Erythematosus Susceptibility. *PLoS Genet*, 7(5):e1002079, 2011. doi: 10.1371/journal.pgen.1002079.

[183] Christopher S Carlson, Michael A Eberle, Mark J Rieder, Qian Yi, Leonid Kruglyak, and Deborah A Nickerson. Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium. *The American Journal of Human Genetics*, 74(1):106–120, 2004. doi: 10.1086/381000.

[184] Lisa L Wang, Kim Worley, Anu Gannavarapu, Murali M Chintagumpala, Moise L Levy, and Sharon E Plon. Intron-Size Constraint as a Mutational Mechanism in Rothmund-Thomson Syndrome. *The American Journal of Human Genetics*, 71(1):165–167, 2002. doi: 10.1086/341234.

[185] R Colobran, E Pedrosa, L Carretero-Iglesia, and M Juan. Copy number variation in chemokine superfamily: the complex scene of CCL3L-CCL4L genes in health and disease. *Clinical & Experimental Immunology*, 162(1):41–52, 2010.

[186] Angela Doring, Christian Gieger, Divya Mehta, Henning Gohlke, Holger Prokisch, Stefan Coassin, Guido Fischer, Kathleen Henke, Norman Klopp, Florian Kronenberg, Bernhard Paulweber, Arne Pfeufer, Dieter Rosskopf, Henry Volzke, Thomas Illig, Thomas Meitinger, H Erich Wichmann, and Christa Meisinger. SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nat Genet*, 40(4):430–436, 2008. doi: 10.1038/ng.107.

[187] Hirotaka Matsuo, Toshinori Chiba, Shushi Nagamori, Akiyoshi Nakayama, Hideharu Domoto, Kanokporn Phetdee, Pattama Wiriyasermkul, Yuichi Kikuchi, Takashi Oda,

Junichiro Nishiyama, Takahiro Nakamura, Yuji Morimoto, Keiko Kamakura, Yutaka Sakurai, Shigeaki Nonoyama, Yoshikatsu Kanai, and Nariyoshi Shinomiya. Mutations in Glucose Transporter 9 Gene SLC2A9 Cause Renal Hypouricemia. *The American Journal of Human Genetics*, 83(6):744–751, 2008. doi: 10.1016/j.ajhg.2008.11.001.

[188] Yen-Ni Teng, Yung-Ming Lin, Ying-Hung Lin, Shu-Yi Tsao, Chao-Chin Hsu, Shio-Jean Lin, Wan-Ching Tsai, and Pao-Lin Kuo. Association of a Single-Nucleotide Polymorphism of the Deleted-in-Azoospermia-Like Gene with Susceptibility to Spermatogenic Failure. *Journal of Clinical Endocrinology & Metabolism*, 87(11):5258–5264, 2002.

[189] Ruslan Dorfman, Andrew Sandford, Chelsea Taylor, Baisong Huang, Daisy Frangolias, Yongqian Wang, Richard Sang, Lilian Pereira, Lei Sun, Yves Berthiaume, Lap-Chee Tsui, Peter D Paré, Peter Durie, Mary Corey, and Julian Zielenski. Complex two-gene modulation of lung disease severity in children with cystic fibrosis. *The Journal of Clinical Investigation*, 118(3):1040–1049, 2008.

[190] Chloe L Thio, Timothy Mosbruger, Jacquie Astemborski, Spencer Greer, Gregory D Kirk, Stephen J O'Brien, and David L Thomas. Mannose Binding Lectin Genotypes Influence Recovery from Hepatitis B Virus Infection. *Journal of Virology*, 79(14): 9192–9196, 2005.

[191] A. Zhang, H. Sun, P. Wang, Y. Han, and X. Wang. Modern analytical techniques in metabolomics analysis. *Analyst.*, 137(2):293–300, 2012. doi: 10.1039/c1an15605e.

[192] S. Collino, F. P. Martin, and S. Rezzi. Clinical metabolomics paves the way towards future healthcare strategies. *Br J Clin Pharmacol.*, 75(3):619–29, 2013. doi: 10.1111/j.1365-2125.2012.04216.x.

[193] HamidM Emwas, RezaM Salek, JulianL Griffin, and Jasmeen Merzaban. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics*, pages 1–25, 2013.

[194] Oliver Fiehn. Metabolomics: the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2):155–171, 2002.

[195] Royston Goodacre. An overflow of what else but metabolism! *Metabolomics*, 6(1): 1–2, 2010.

[196] Laeticia Da Silva, Markus Godejohann, François-Pierre J. Martin, Sebastiano Collino, Alexander Bürkle, María Moreno-Villanueva, Jörgen Bernhardt, Olivier Toussaint, Beatrix Grubeck-Loebenstein, Efstathios S. Gonos, Ewa Sikora, Tilman Grune, Nicolle Breusing, Claudio Franceschi, Antti Hervonen, Manfred Spraul, and Sofia Moco. High-resolution quantitative metabolome analysis of urine by automated flow injection NMR. *Analytical Chemistry*, 85(12):5801–5809, 2013.

[197] Michael Lauridsen, Steen H. Hansen, Jerzy W. Jaroszewski, and Claus Cornett. Human urine as test material in 1H NMR-based metabonomics: Recommendations for sample preparation and storage. *Analytical Chemistry*, 79(3):1181–1186, 2007.

[198] Huifeng Wu, Andrew D. Southam, Adam Hines, and Mark R. Viant. High-throughput tissue extraction protocol for NMR- and ms-based metabolomics. *Analytical Biochemistry*, 372(2):204–212, 2008.

[199] Manuel Liebeke, Jie Hao, Timothy M. D. Ebbels, and Jacob G. Bundy. Combining spectral ordering with peak fitting for one-dimensional NMR quantitative metabolomics. *Analytical Chemistry*, 85(9):4605–4612, 2013.

[200] Neil MacKinnon, Wencheng Ge, Amjad P. Khan, Bagganahalli S. Somashekar, Pratima Tripathi, Javed Siddiqui, John T. Wei, Arul M. Chinnaiyan, Thekkelnaycke M. Rajendiran, and Ayyalusamy Ramamoorthy. Variable reference alignment: An improved peak alignment protocol for NMR spectral data with large intersample varia-

tion. *Analytical Chemistry*, 84(12):5372–5379, 2012.

[201] Neil MacKinnon, Bagganahalli S. Somashekar, Pratima Tripathi, Wencheng Ge, Thekkelnaycke M. Rajendiran, Arul M. Chinnaiyan, and Ayyalusamy Ramamoorthy. Metaboid: A graphical user interface package for assignment of 1H NMR spectra of bodyfluids and tissues. *Journal of Magnetic Resonance*, 226(0):93–99, 2013.

[202] Agnieszka Smolinska, Lionel Blanchet, Lutgarde M. C. Buydens, and Sybren S. Wijmenga. NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta*, 750(0):82–97, 2012.

[203] S. Zhang, G. A. Nagana Gowda, T. Ye, and D. Raftery. Advances in NMR-based biofluid analysis and metabolite profiling. *Analyst.*, 135(7):1490–8, 2010. doi: 10.1039/c000091d.

[204] D. F. Ransohoff. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer.*, 5(2):142–9., 2005.

[205] Royston Goodacre, David Broadhurst, AgeK Smilde, BruceS Kristal, J. David Baker, Richard Beger, Conrad Bessant, Susan Connor, Giorgio Capuani, Andrew Craig, Tim Ebbels, DouglasB Kell, Cesare Manetti, Jack Newton, Giovanni Paternostro, Ray Somorjai, Michael Sjöström, Johan Trygg, and Florian Wulfert. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3(3):231–241, 2007.

[206] Athanassios Kassidas, John F. MacGregor, and Paul A. Taylor. Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, 44(4):864–875, 1998.

[207] Antonio Julià, Eugeni Domènech, Elena Ricart, Raül Tortosa, Valle García-Sánchez, Javier Pérez Gisbert, Pilar Nos Mateu, Ana Gutiérrez, Fernando Gomollón, Juan Luís Mendoza, Esther Garcia-Planella, Manuel Barreiro-de Acosta, Fernando Muñoz, Maribel Vera, Cristina Saro, Maria Esteve, Montserrat Andreu, Arnald Alonso, María López-Lasanta, Laia Codó, Josep Lluís Gelpí, Andres C. García-Montero, Jaume Bertranpetit, Devin Absher, Julián Panés, and Sara Marsal. A genome-wide association study on a southern european population identifies a new Crohn's disease susceptibility locus at RBX1-EP300. *Gut*, 2012.

[208] Jenny Forshed, Ralf J. O. Torgrip, K. Magnus a?berg, Bo Karlberg, Johan Lindberg, and Sven P. Jacobsson. A comparison of methods for alignment of NMR peaks in the context of cluster analysis. *Journal of Pharmaceutical and Biomedical Analysis*, 38(5):824–832, 2005.

[209] R. J. Xavier and D. K. Podolsky. Unravelling the pathogenesis of inflammatory bowel disease. *Nature*, 448(7152):427–434, 2007.

[210] Luke Jostins, Stephan Ripke, Rinse K. Weersma, Richard H. Duerr, Dermot P. McGovern, Ken Y. Hui, James C. Lee, L. Philip Schumm, Yashoda Sharma, Carl A. Anderson, Jonah Essers, Mitja Mitrovic, Kaida Ning, Isabelle Cleynen, Emilie Theatre, Sarah L. Spain, Soumya Raychaudhuri, Philippe Goyette, Zhi Wei, Clara Abraham, Jean-Paul Achkar, Tariq Ahmad, Leila Amininejad, Ashwin N. Ananthakrishnan, Vibeke Andersen, Jane M. Andrews, Leonard Baidoo, Tobias Balschun, Peter A. Bampton, Alain Bitton, Gabrielle Boucher, Stephan Brand, Carsten Buning, Ariella Cohain, Sven Cichon, Mauro D/'Amato, Dirk De Jong, Kathy L. Devaney, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Lynnette R. Ferguson, Denis Franchimont, Karin Fransen, Richard Gearry, Michel Georges, Christian Gieger, Jurgen Glas, Talin Haritunians, Ailsa Hart, Chris Hawkey, Matija Hedl, Xinli Hu, Tom H. Karlsen, Limas Kupcinskas, Subra Kugathasan, Anna Latiano, Debby Laukens, Ian C. Lawrance, Charlie W. Lees, Edouard Louis, Gillian Mahy, John Mansfield, Angharad R. Morgan, Craig Mowat,

William Newman, Orazio Palmieri, Cyriel Y. Ponsioen, Uros Potocnik, Natalie J. Prescott, Miguel Regueiro, Jerome I. Rotter, Richard K. Russell, Jeremy D. Sanderson, Miquel Sans, Jack Satsangi, Stefan Schreiber, Lisa A. Simms, Jurgita Sventoraityte, Stephan R. Targan, Kent D. Taylor, Mark Tremelling, Hein W. Verspaget, Martine De Vos, Cisca Wijmenga, David C. Wilson, Juliane Winkelmann, Ramnik J. Xavier, Sebastian Zeissig, Bin Zhang, Clarence K. Zhang, Hongyu Zhao, Mark S. Silverberg, Vito Annese, Hakon Hakonarson, Steven R. Brant, Graham Radford-Smith, Christopher G. Mathew, John D. Rioux, Eric E. Schadt, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422): 119–124, 2012.

[211] Bernard Khor, Agnes Gardet, and Ramnik J. Xavier. Genetics and pathogenesis of inflammatory bowel disease. *Nature*, 474(7351):307–317, 2011.

[212] J. E. Lennard-Jones. Classification of inflammatory bowel disease. *Scand J Gastroenterol Suppl.*, 170:2–6; discussion 16–9., 1989.

[213] M. Chaparro, J. Panes, V. Garcia, M. Manosa, M. Esteve, O. Merino, M. Andreu, A. Gutierrez, F. Gomollon, J. L. Cabriada, M. A. Montoro, J. L. Mendoza, P. Nos, and J. P. Gisbert. Long-term durability of infliximab treatment in Crohn's disease and efficacy of dose "escalation" in patients losing response. *J Clin Gastroenterol.*, 45(2): 113–8, 2011. doi: 10.1097/MCG.0b013e3181ebaef9.

[214] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–909, 2006.

[215] Olivier Delaneau, Jean-Francois Zagury, and Jonathan Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Meth*, 10(1): 5–6, 2013.

[216] B. Howie, J. Marchini, and M. Stephens. Genotype imputation with thousands of genomes. *G3 (Bethesda).*, 1(6):457–70, 2011. doi: 10.1534/g3.111.001198.

[217] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[218] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[219] Mark I. McCarthy, Goncalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369, 2008.

[220] Manolis Kellis, Barbara Wold, Michael P. Snyder, Bradley E. Bernstein, Anshul Kundaje, Georgi K. Marinov, Lucas D. Ward, Ewan Birney, Gregory E. Crawford, Job Dekker, Ian Dunham, Laura L. Elnitski, Peggy J. Farnham, Elise A. Feingold, Mark Gerstein, Morgan C. Giddings, David M. Gilbert, Thomas R. Gingeras, Eric D. Green, Roderic Guigo, Tim Hubbard, Jim Kent, Jason D. Lieb, Richard M. Myers, Michael J. Pazin, Bing Ren, John A. Stamatoyannopoulos, Zhiping Weng, Kevin P. White, and Ross C. Hardison. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17):6131–6138, 2014.

[221] Boyko Kabakchiev and Mark S. Silverberg. Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine.

Gastroenterology, 144(7):1488–1496.e3, 2013.

[222] Terry M Therneau. Modeling survival data: extending the cox model. *Springer*, 2000.

[223] R Development Core Team. R: A Language and Environment for Statistical Computing. 2011.

[224] Augustin Luna and Kristin K. Nicodemus. snp.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package. *Bioinformatics*, 23(6):774–776, 2007.

[225] M. J. Koslowski, I. Kubler, M. Chamaillard, E. Schaeffeler, W. Reinisch, G. Wang, J. Beisner, A. Teml, L. Peyrin-Biroulet, S. Winter, K. R. Herrlinger, P. Rutgeerts, S. Vermeire, R. Cooney, K. Fellermann, D. Jewell, C. L. Bevins, M. Schwab, E. F. Stange, and J. Wehkamp. Genetic variants of wnt transcription factor TCF-4 (TCF7L2) putative promoter region are associated with small intestinal Crohn's disease. *PLoS One.*, 4 (2):e4496, 2009. doi: 10.1371/journal.pone.0004496.

[226] Julia Dambacher, Tanja Staudinger, Julia Seiderer, Zeljka Sisic, Fabian Schnitzler, Simone Pfennig, Katrin Hofbauer, Astrid Konrad, Cornelia Tillack, Jan-Michel Otte, Joachim Diebold, Burkhard Göke, Thomas Ochsenkühn, Peter Lohse, and Stephan Brand. Macrophage migration inhibitory factor (MIF) -173G/C promoter polymorphism influences upper gastrointestinal tract involvement and disease activity in patients with Crohn's disease. *Inflammatory Bowel Diseases*, 13(1):71–82, 2007.

[227] Sharyle A. Fowler, Ashwin N. Ananthakrishnan, Agnes Gardet, Christine R. Stevens, Joshua R. Korzenik, Bruce E. Sands, Mark J. Daly, Ramnik J. Xavier, and Vijay Yajnik. SMAD3 gene variant is a risk factor for recurrent surgery in patients with Crohn's disease. *Journal of Crohn's and Colitis*, 8(8):845–851, 2014.

[228] V. Marcil, D. Sinnett, E. Seidman, F. Boudreau, F. P. Gendron, J. F. Beaulieu, D. Menard, M. Lambert, A. Bitton, R. Sanchez, D. Amre, and E. Levy. Association between genetic variants in the HNF4A gene and childhood-onset Crohn's disease. *Genes Immun*, 13 (7):556–565, 2012.

[229] Tanja Zeller, Philipp Wild, Silke Szymczak, Maxime Rotival, Arne Schillert, Raphaele Castagne, Seraya Maouche, Marine Germain, Karl Lackner, Heidi Rossmann, Medea Eleftheriadis, Christoph R. Sinning, Renate B. Schnabel, Edith Lubos, Detlev Mennerich, Werner Rust, Claire Perret, Carole Proust, Viviane Nicaud, Joseph Loscalzo, Norbert Hübner, David Tregouet, Thomas Münzel, Andreas Ziegler, Laurence Tiret, Stefan Blankenberg, and François Cambien. Genetics and beyond - the transcriptome of human monocytes and disease susceptibility. *PLoS ONE*, 5(5):e10693, 2010.

[230] I. Dobrosotskaya, R. K. Guy, and G. L. James. MAGI-1, a membrane-associated guanylate kinase with a unique arrangement of protein-protein interaction domains. *J Biol Chem.*, 272(50):31589–97., 1997.

[231] S. Hirabayashi, M. Tajima, I. Yao, W. Nishimura, H. Mori, and Y. Hata. JAM4, a junctional cell adhesion molecule interacting with a tight junction protein, MAGI-1. *Mol Cell Biol.*, 23(12):4267–82, 2003.

[232] Richard P. Laura, Sarajane Ross, Hartmut Koeppen, and Laurence A. Lasky. MAGI-1: A widely expressed, alternatively spliced tight junction protein. *Experimental Cell Research*, 275(2):155–170, 2002.

[233] L. Shen, C. R. Weber, D. R. Raleigh, D. Yu, and J. R. Turner. Tight junction pore and leak pathways: a dynamic duo. *Annu Rev Physiol.*, 73:283–309, 2011. doi: 10.1146/annurev-physiol-012110-142150.

[234] Makiko Tajima, Susumu Hirabayashi, Ikuko Yao, Madoka Shirasawa, Junichi Osuga, Shun Ishibashi, Toshiro Fujita, and Yutaka Hata. Roles of immunoglobulin-like loops

of junctional cell adhesion molecule 4; involvement in the subcellular localization and the cell adhesion. *Genes to Cells*, 8(9):759–768, 2003.

[235] S. Garrido-Urbani, P. F. Bradfield, and B. A. Imhof. Tight junction dynamics: the role of junctional adhesion molecules (JAMs). *Cell and Tissue Research*, 355(3):701–715, 2014.

[236] P. Henderson, J. E. van Limbergen, J. Schwarze, and D. C. Wilson. Function of the intestinal epithelium and its dysregulation in inflammatory bowel disease. *Inflamm Bowel Dis.*, 17(1):382–95, 2011. doi: 10.1002/ibd.21379.

[237] Amanda M. Marchiando, W. Vallen Graham, and Jerrold R. Turner. Epithelial barriers in homeostasis and disease. *Annual Review of Pathology: Mechanisms of Disease*, 5 (1):119–144, 2010.

[238] J. R. Turner. Intestinal mucosal barrier function in health and disease. *Nat Rev Immunol.*, 9(11):799–809, 2009. doi: 10.1038/nri2653.

[239] D. R. Clayburgh, L. Shen, and J. R. Turner. A porous defense: the leaky epithelial barrier in intestinal disease. *Lab Invest.*, 84(3):282–91., 2004.

[240] M. A. McGuckin, R. Eri, L. A. Simms, T. H. Florin, and G. Radford-Smith. Intestinal barrier dysfunction in inflammatory bowel diseases. *Inflamm Bowel Dis.*, 15(1):100–13, 2009. doi: 10.1002/ibd.20539.

[241] Norimasa Sawada. Tight junction-related human diseases. *Pathology International*, 63(1):1–12, 2013.

[242] J. Benjamin, G. K. Makharia, V. Ahuja, M. Kalaivani, and Y. K. Joshi. Intestinal permeability and its association with the patient and disease characteristics in Crohn's disease. *World J Gastroenterol.*, 14(9):1399–405., 2008.

[243] N. Gassler, C. Rohr, A. Schneider, J. Kartenbeck, A. Bach, N. Obermuller, H. F. Otto, and F. Autschbach. Inflammatory bowel disease is associated with changes of enterocytic junctions. *Am J Physiol Gastrointest Liver Physiol.*, 281(1):G216–28., 2001.

[244] M. Peeters, B. Geypens, D. Claus, H. Nevens, Y. Ghoos, G. Verbeke, F. Baert, S. Vermeire, R. Vlietinck, and P. Rutgeerts. Clustering of increased small intestinal permeability in families with Crohn's disease. *Gastroenterology.*, 113(3):802–7., 1997.

[245] Anny-Claude Luissint, Asma Nusrat, and CharlesA Parkos. JAM-related proteins in mucosal homeostasis and inflammation. *Seminars in Immunopathology*, pages 1–16, 2014.

[246] Lance W. Peterson and David Artis. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nat Rev Immunol*, 14(3):141–153, 2014.

[247] Atle van Beelen Granlund, Arnar Flatberg, Ann E. Ostvik, Ignat Drozdov, Bjørn I. Gustafsson, Mark Kidd, Vidar Beisvag, Sverre H. Torp, Helge L. Waldum, Tom Christian Martinsen, Jan Kristian Damas, Terje Espevik, and Arne K. Sandvik. Whole genome gene expression meta-analysis of inflammatory bowel disease colon mucosa demonstrates lack of major differences between Crohn's disease and ulcerative colitis. *PLoS ONE*, 8(2):e56818, 2013.

[248] Talin Haritunians, Kent D. Taylor, Stephan R. Targan, Marla Dubinsky, Andrew Ippoliti, Soonil Kwon, Xiuqing Guo, Gil Y. Melmed, Dror Berel, Emebet Mengesha, Bruce M. Psaty, Nicole L. Glazer, Eric A. Vasiliauskas, Jerome I. Rotter, Phillip R. Fleshner, and Dermot P. B. McGovern. Genetic predictors of medically refractory ulcerative colitis. *Inflammatory Bowel Diseases*, 16(11):1830–1840, 2010.

[249] A. Jauregi-Miguel, N. Fernandez-Jimenez, I. Irastorza, L. Plaza-Izurieta, J. C. Vitoria, and J. R. Bilbao. Alteration of tight junction gene expression in celiac disease. *J Pediatr Gastroenterol Nutr*, 58(6):767–7, 2014.

[250] E. Norén, S. Almer, and J. Söderman. Association between genetic markers related to tight junctions and inflammatory bowel disease. *9th Congress of ECCO*, P664, 2014.

[251] A. Kobayashi, D. S. Donaldson, T. Kanaya, S. Fukuda, J. K. Baillie, T. C. Freeman, H. Ohno, I. R. Williams, and N. A. Mabbott. Identification of novel genes selectively expressed in the follicle-associated epithelium from the meta-analysis of transcriptomics data from multiple mouse cell and tissue populations. *DNA Res.*, 19(5):407–22, 2012. doi: 10.1093/dnares/dss022.

[252] N. A. Mabbott, D. S. Donaldson, H. Ohno, I. R. Williams, and A. Mahajan. Microfold (m) cells: important immunosurveillance posts in the intestinal epithelium. *Mucosal Immunol*, 6(4):666–677, 2013.

[253] S. A. Bustin, S. R. Li, and S. Dorudi. Expression of the Ca2+-activated Chloride Channel Genes CLCA1 and CLCA2 is Downregulated in Human Colorectal Cancer. *DNA Cell Biol.*, 20(6):331–8., 2001.

[254] Rikako Suzuki, Shingo Miyamoto, Yumiko Yasui, Shigeyuki Sugie, and Takuji Tanaka. Global gene expression analysis of the mouse colonic mucosa treated with azoxymethane and dextran sodium sulfate. *BMC Cancer*, 7(1):84, 2007.

[255] P. H. Schafer, A. Parton, A. K. Gandhi, L. Capone, M. Adams, L. Wu, J. B. Bartlett, M. A. Loveland, A. Gilhar, Y. F. Cheung, G. S. Baillie, M. D. Houslay, H. W. Man, G. W. Muller, and D. I. Stirling. Apremilast, a camp phosphodiesterase-4 inhibitor, demonstrates anti-inflammatory activity in vitro and in a model of psoriasis. *British Journal of Pharmacology*, 159(4):842–855, 2010.

[256] Pooneh Salari and Mohammad Abdollahi. Phosphodiesterase inhibitors in inflammatory bowel disease. *Expert Opinion on Investigational Drugs*, 21(3):261–264, 2012.

[257] R. E. Shrimpton, M. Butler, A. S. Morel, E. Eren, S. S. Hue, and M. A. Ritter. CD205 (dec-205): a recognition receptor for apoptotic and necrotic self. *Mol Immunol.*, 46 (6):1229–39, 2009. doi: 10.1016/j.molimm.2008.11.016.

[258] Kayo Inaba, William J. Swiggard, Muneo Inaba, Joseph Meltzer, Asra Miryza, Tatsuya Sasagawa, Michel C. Nussenzweig, and Ralph U. Steinman. Tissue distribution of the DEC-205 protein that is detected by the monoclonal antibody NLDC-145: I. expression on dendritic cells and other subsets of mouse leukocytes. *Cellular Immunology*, 163(1):148–156, 1995.

[259] Jurjen Tel, Daniel Benitez-Ribas, Sander Hoosemans, Alessandra Cambi, Gosse J. Adema, Carl G. Figdor, Paul J. Tacken, and I. Jolanda M. de Vries. Dec-205 mediates antigen uptake and presentation by both resting and activated human plasmacytoid dendritic cells. *European Journal of Immunology*, 41(4):1014–1023, 2011.

[260] Ryan E Mills, Klaudia Walter, Chip Stewart, Robert E Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtai Chris Yoon, Kai Ye, R Keira Cheetham, Asif Chinwalla, Donald F Conrad, Yutao Fu, Fabian Grubert, Iman Hajirasouliha, Fereydoun Hormozdiari, Lilia M Iakoucheva, Zamin Iqbal, Shuli Kang, Jeffrey M Kidd, Miriam K Konkel, Joshua Korn, Ekta Khurana, Deniz Kural, Hugo Y K Lam, Jing Leng, Ruiqiang Li, Yingrui Li, Chang-Yun Lin, Ruibang Luo, Xinmeng Jasmine Mu, James Nemesh, Heather E Peckham, Tobias Rausch, Aylwyn Scally, Xinghua Shi, Michael P Stromberg, Adrian M Stutz, Alexander Eckehart Urban, Jerilyn A Walker, Jiantao Wu, Yujun Zhang, Zhengdong D Zhang, Mark A Batzer, Li Ding, Gabor T Marth, Gil McVean, Jonathan Sebat, Michael Snyder, Jun Wang, Kenny Ye, Evan E Eichler, Mark B Gerstein, Matthew E Hurles, Charles Lee, Steven A McCarroll, and Jan O Korbel. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, 2011. doi: 10.1038/nature09708.

[261] Antonio Julià, José Antonio Pinto, Jordi Gratacós, Rubén Queiró, Carlos Ferrándiz, Eduardo Fonseca, Carlos Montilla, Juan Carlos Torre-Alonso, Lluís Puig, José Javier Pérez Venegas, et al. A deletion at ADAMTS9-MAGI1 locus is associated with psoriatic arthritis risk. *Annals of the rheumatic diseases*, pages annrheumdis–2014, 2015.

[262] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

[263] J. R. Gibbs, M. P. van der Brug, D. G. Hernandez, B. J. Traynor, M. A. Nalls, S. L. Lai, S. Arepalli, A. Dillman, I. P. Rafferty, J. Troncoso, R. Johnson, H. R. Zielke, L. Ferrucci, D. L. Longo, M. R. Cookson, and A. B. Singleton. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.*, 6(5): e1000952, 2010. doi: 10.1371/journal.pgen.1000952.

[264] E. Grundberg, E. Meduri, J. K. Sandling, A. K. Hedman, S. Keildson, A. Buil, S. Busche, W. Yuan, J. Nisbet, M. Sekowska, A. Wilk, A. Barrett, K. S. Small, B. Ge, M. Caron, S. Y. Shin, M. Lathrop, E. T. Dermitzakis, M. I. McCarthy, T. D. Spector, J. T. Bell, and P. Deloukas. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet.*, 93(5):876–90, 2013. doi: 10.1016/j.ajhg.2013.10.004.

[265] E. Grundberg, K. S. Small, A. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T. P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S. Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, and T. D. Spector. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet.*, 44(10):1084–9, 2012. doi: 10.1038/ng.2394.

[266] T. Lappalainen, M. Sammeth, M. R. Friedlander, P. A. t Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlof, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Hasler, A. C. Syvanen, G. J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigo, I. G. Gut, X. Estivill, and E. T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature.*, 501(7468): 506–11, 2013. doi: 10.1038/nature12531.

[267] S. B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature.*, 464(7289):773–7, 2010. doi: 10.1038/nature08903.

[268] E. E. Schadt, C. Molony, E. Chudin, K. Hao, X. Yang, P. Y. Lum, A. Kasarskis, B. Zhang, S. Wang, C. Suver, J. Zhu, J. Millstein, S. Sieberts, J. Lamb, D. GuhaThakurta, J. Derry, J. D. Storey, I. Avila-Campillo, M. J. Kruger, J. M. Johnson, C. A. Rohl, A. van Nas, M. Mehrabian, T. A. Drake, A. J. Lusis, R. C. Smith, F. P. Guengerich, S. C. Strom,

E. Schuetz, T. H. Rushmore, and R. Ulrich. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, 6(5):e107, 2008. doi: 10.1371/journal.pbio.0060107.

[269] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, 306(5696):636–640, 2004.

[270] Kathi Canese and Sarah Weis. Pubmed: The bibliographic database. 2013.

[271] C. L. Noble, A. R. Abbas, J. Cornelius, C. W. Lees, G. T. Ho, K. Toy, Z. Modrusan, N. Pal, F. Zhong, S. Chalasani, H. Clark, I. D. Arnott, I. D. Penman, J. Satsangi, and L. Diehl. Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. *Gut.*, 57(10):1398–405, 2008. doi: 10.1136/gut.2008.148395.

[272] C. L. Noble, A. R. Abbas, C. W. Lees, J. Cornelius, K. Toy, Z. Modrusan, H. F. Clark, I. D. Arnott, I. D. Penman, J. Satsangi, and L. Diehl. Characterization of intestinal gene expression profiles in Crohn's disease by genome-wide microarray analysis. *Inflamm Bowel Dis.*, 16(10):1717–28, 2010. doi: 10.1002/ibd.21263.

[273] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, 2014.

[274] Hongsheng Gui, Miaoxin Li, Pak Sham, and Stacey Cherny. Comparisons of seven algorithms for pathway analysis using the WTCCC Crohn's disease dataset. *BMC Research Notes*, 4(1):386, 2011. ISSN 1756-0500.

A | Supplementary Data of GStream

A.1 Supplementary figures

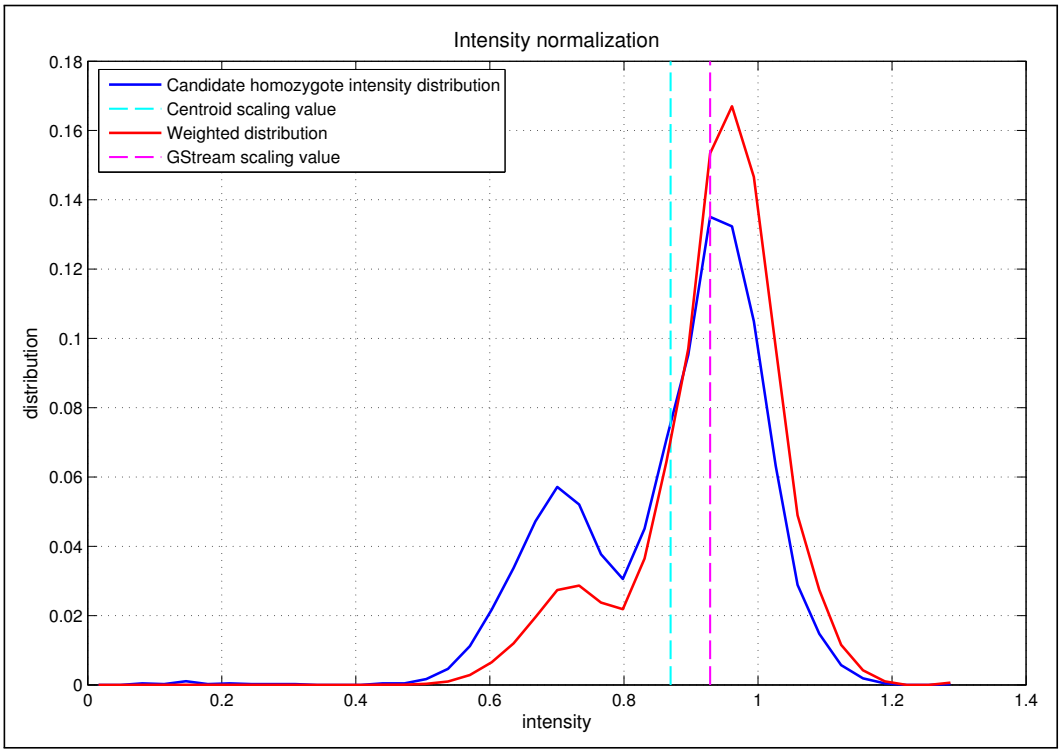


Figure A.1: *Example of raw intensity normalization.* The intensity distribution of candidate homozygote samples (i.e. AA) across its specific allele channel (i.e. channel A) is plotted together with its centroid scaling value as computed by Peiffer et al.[75]. GStream first weights this distribution and computes its maximum to scale the channel intensities by the corresponding intensity value. This example shows a typical CNV pattern where the error produced by the first approach is magnified.

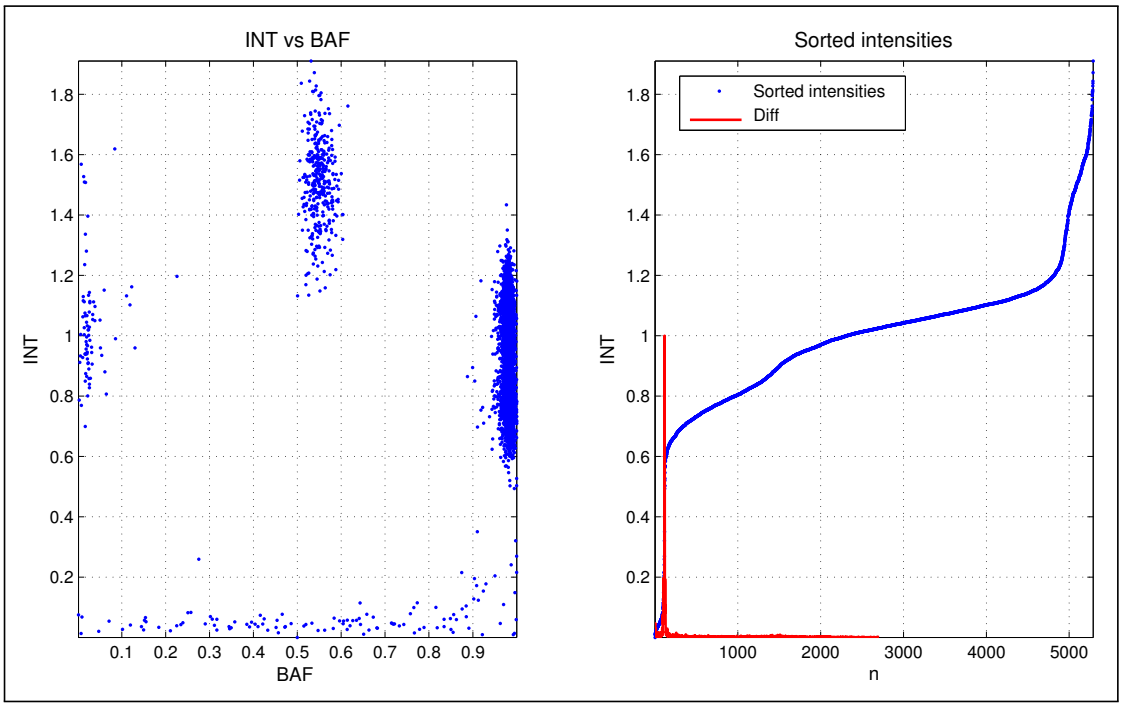


Figure A.2: *Example of how zero-threshold is computed.* (A) BAF and absolute intensities of an example marker where some homozygous deletion samples with low intensity values can be observed. (B) Absolute intensities are sorted and differences between consecutive sorted intensities normalized to one. The observed peak over these differences points to the intensity value that will be set as threshold.

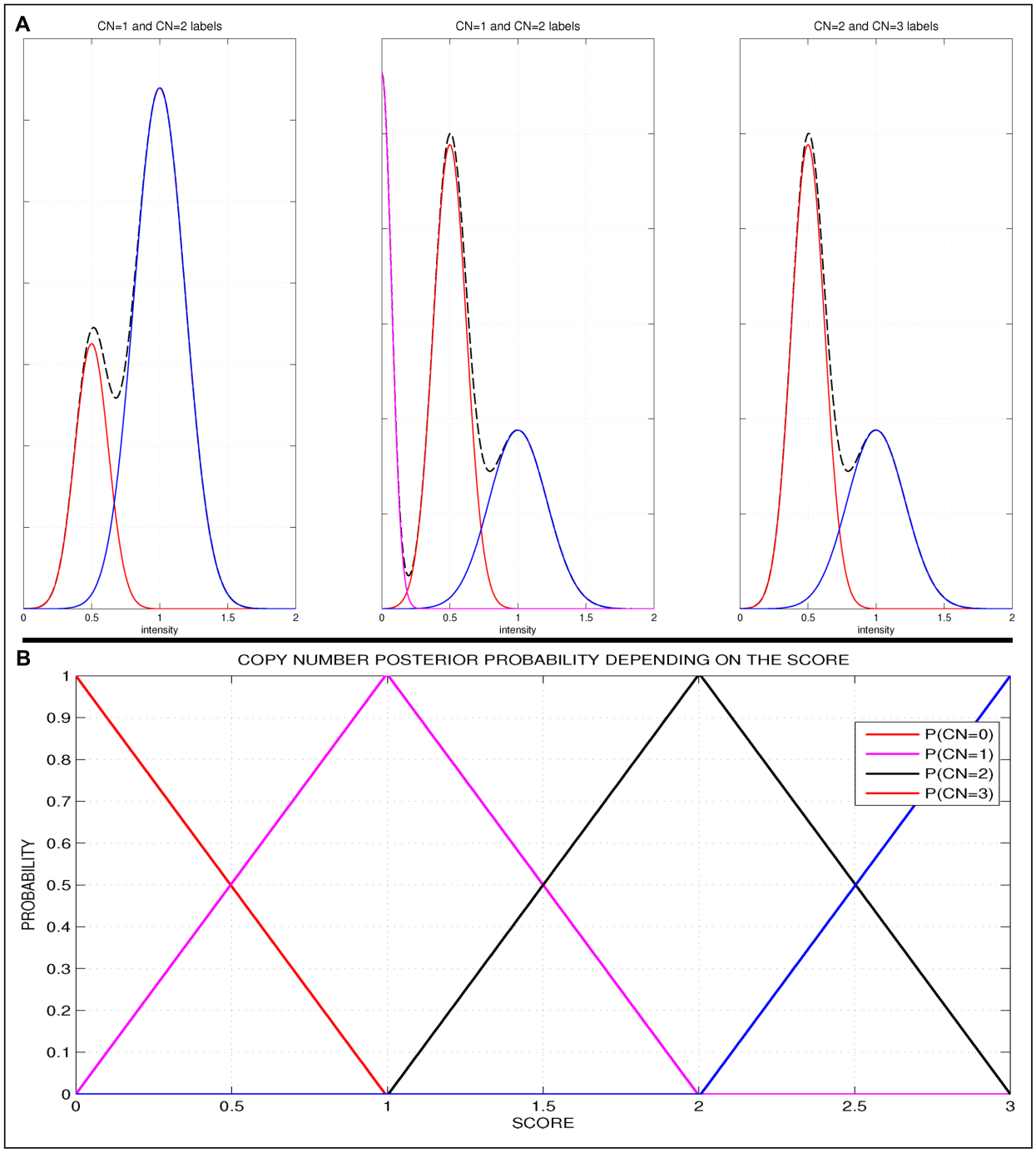


Figure A.3: *CNV labelling and scoring.* (A). Category disambiguation when the two-component model is selected. The leftmost graph shows a case where the higher intensity component (blue) is more frequent and it is assigned to the diploid state while the lower intensity component (red) is assigned to the deletion state. This assignment is due to the fact that high frequency amplifications are very uncommon and undetectable with this technology. The centre graph shows a case where the higher intensity component is less frequent and homozygous deletion samples have been found (magenta). In this case, the higher component (blue) is assigned to the diploid state and the lower component (red) to the deletion state fulfilling the expected Hardy-Weinberg equilibrium frequencies. Finally, the rightmost graph shows the last case where the higher intensity component is less frequent and no homozygous deletion samples have been found. In this case the higher component is assigned to the amplification state and the lower component to the diploid state. (B) Posterior probability of each copy number depending on the score assigned by GStream: From 0 to 0.5 samples can be categorized as homozygous deletion, from 0.5 to 1.5 as hemizygous deletion, from 1.5 to 2.5 as diploid and from 2.5 to 3 as amplification.

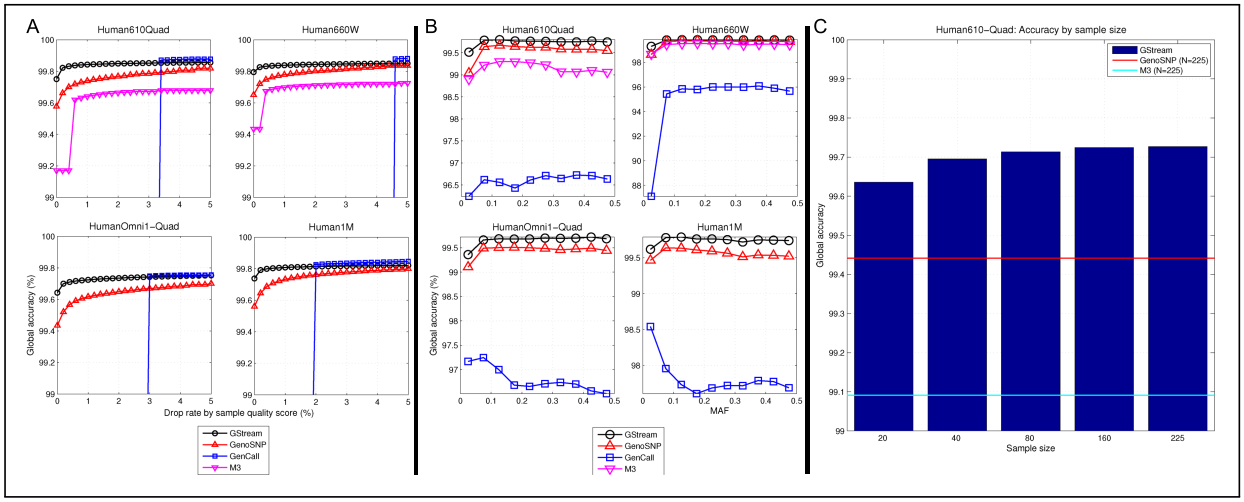


Figure A.4: *Genotyping performance.* (A) Genotyping performance depending on the drop rate, where calls dropped from the accuracy analysis were selected according to the genotype call quality score. (B) Genotyping performance depending on the SNP minor allele frequency. (C) Genotyping accuracy of GStream at different sample sizes (i.e. N = 20, 40, 60, 80, 160 and 225) compared to the accuracies obtained by GenoSNP and M3 with the highest sample size (N = 225).

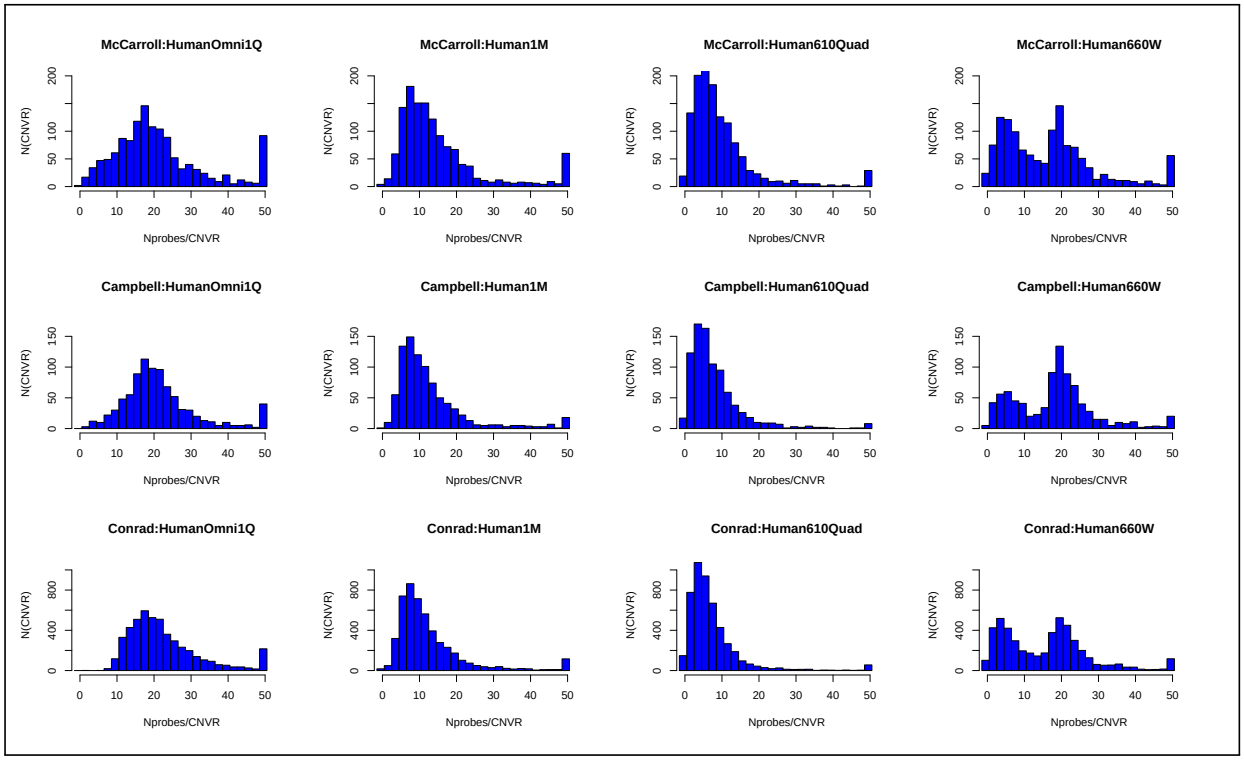


Figure A.5: *Microarray coverage density.* Coverage density of each microarray platform over the CNV regions defined by each reference study. There are major differences between the first Infinum HD platforms (Human610-Quad and Human1M-Duo) and their succesors including specific CNV coverage (Human660W-Quad and HumanOmni1-Quad). Both Human610-Quad and Human1M-Duo have a mean number of 10 markers covering CNV regions, while Human660W was designed with a highest coverage (20 markers/region) for almost 50% of the regions. Finally, HumanOmni1-Quad increased the global coverage to 20 markers/region.

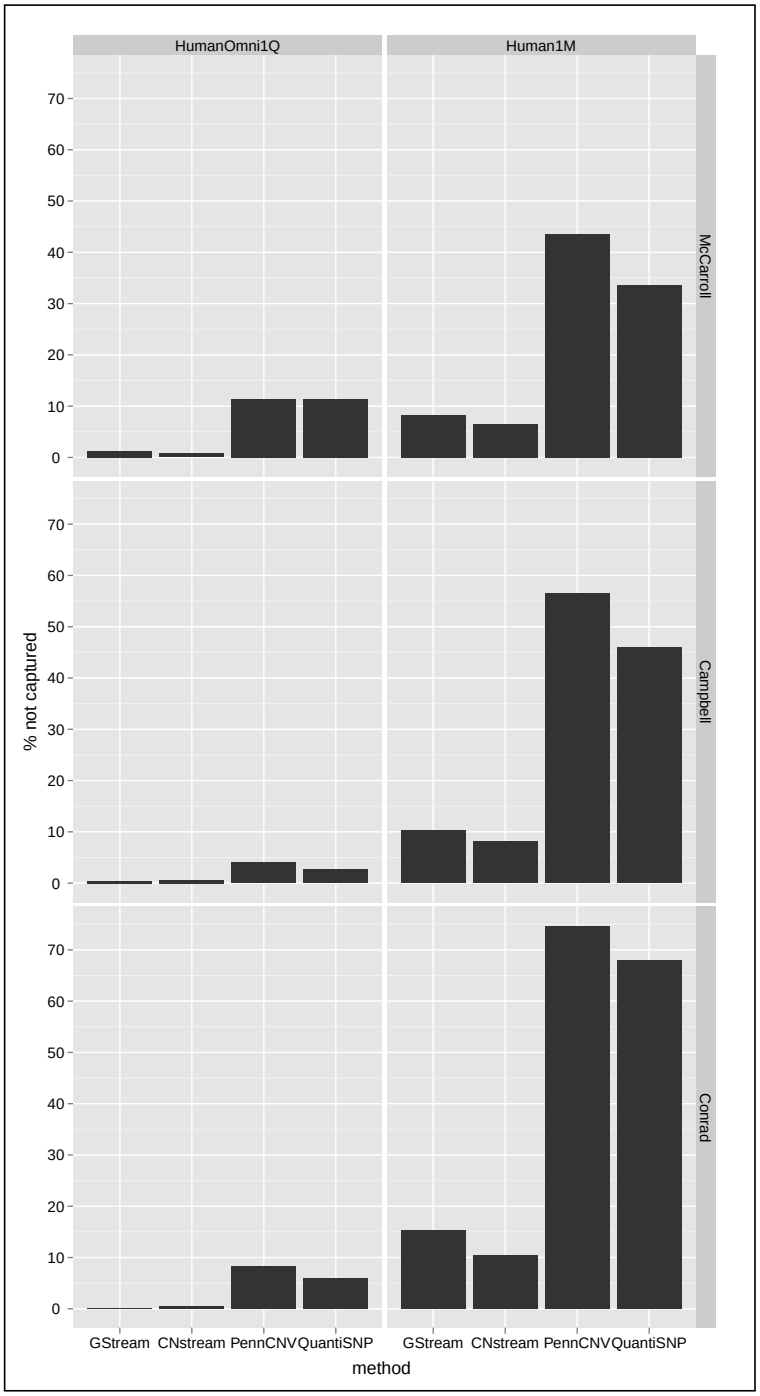


Figure A.6: *Missed associations.* Percentage of associations (i.e. *P*-Value<0.05 over the golden standard dataset) that were not captured by the different methods tested (i.e. *P*-Value>0.05 over the tested method).

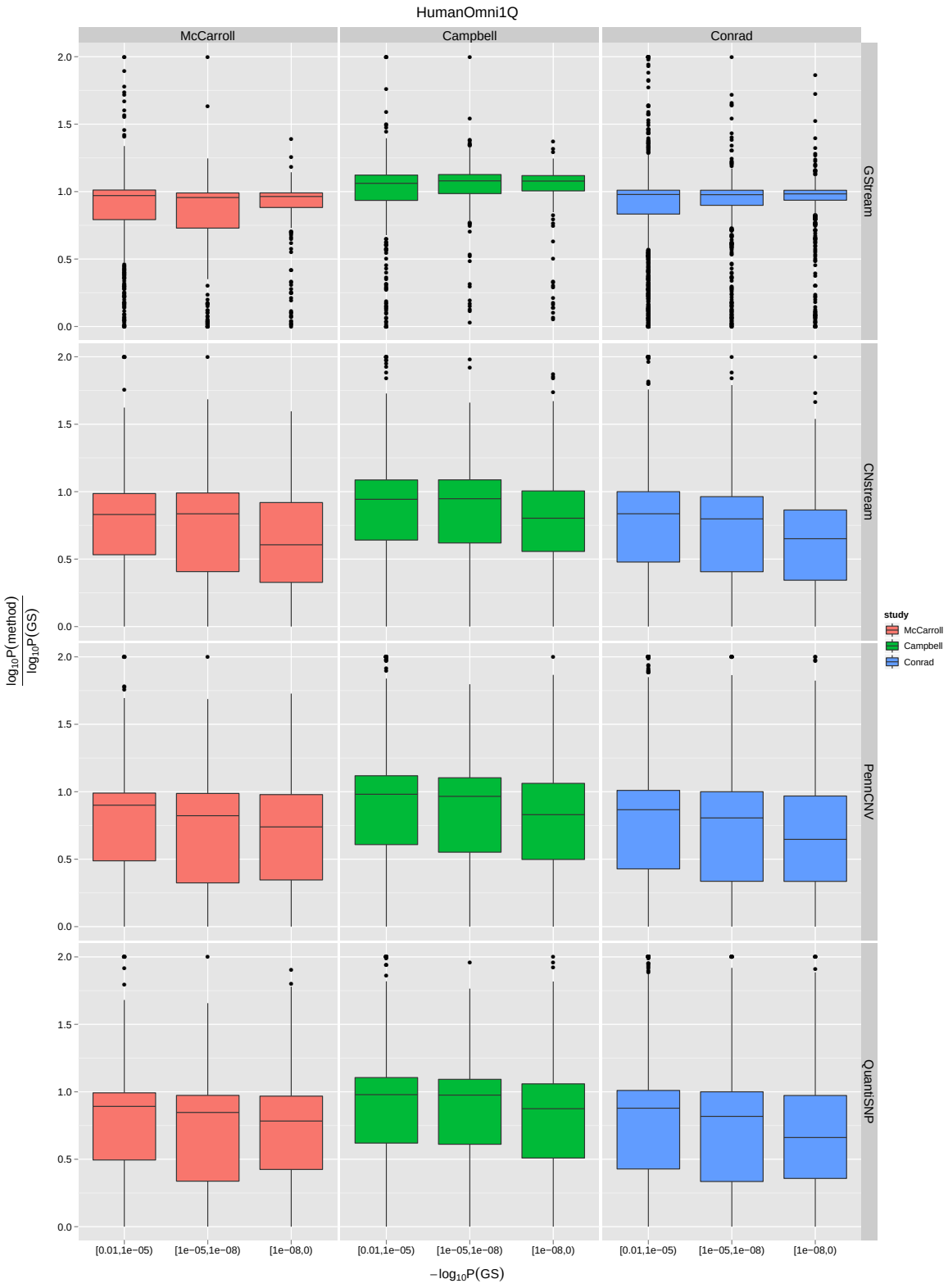


Figure A.7: *HumanOmni1-Quad P-Value distributions.* Distributions of the *P*-Value association ratios depending on the golden standard dataset used for evaluation (i.e. represented by different colours) and on the *P*-Value range obtained over the golden standard calls (i.e. horizontal axis).

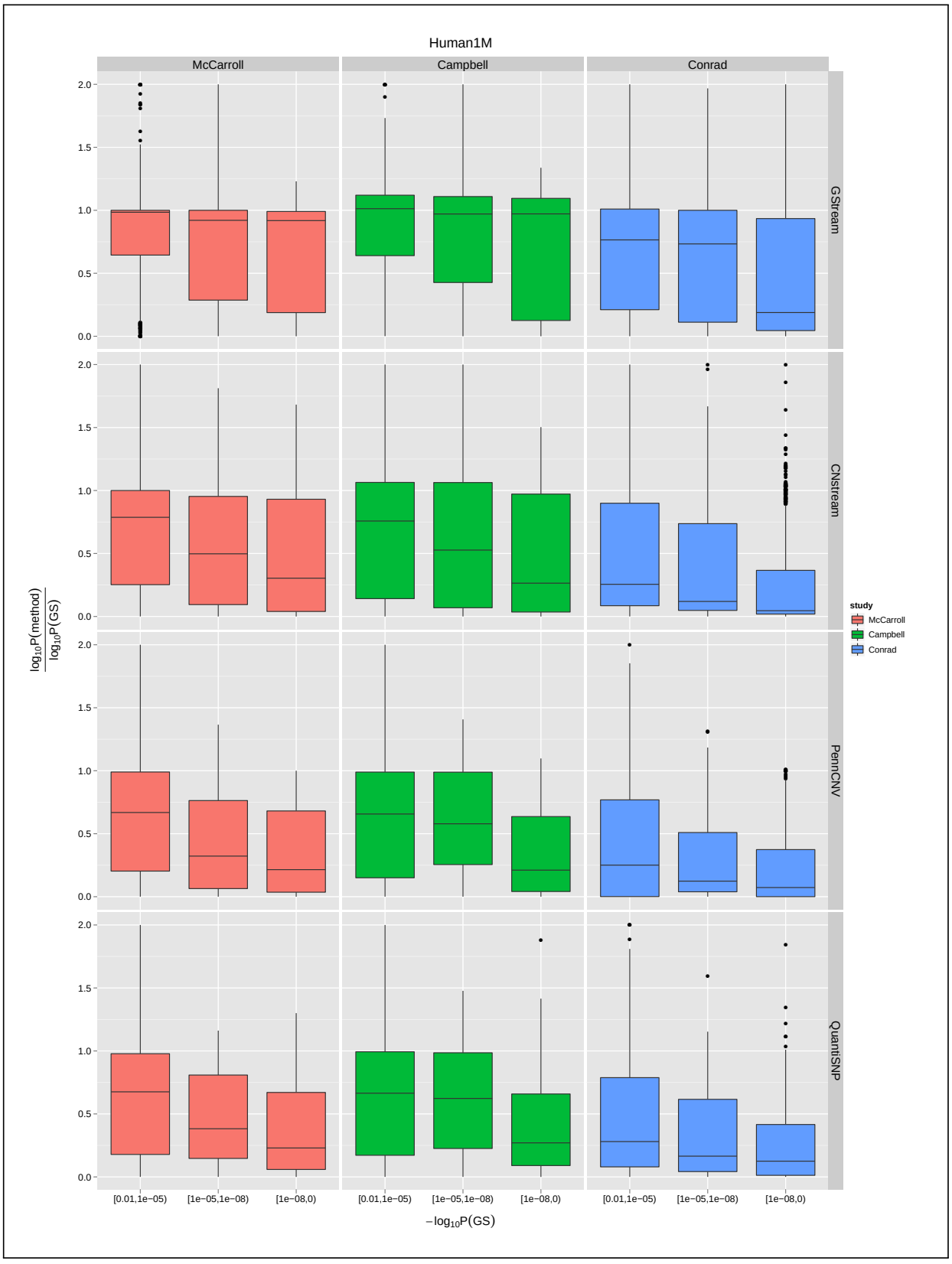


Figure A.8: *1M-Duo P-Value distributions.* Distributions of the *P*-Value association ratios depending on the golden standard dataset used for evaluation (i.e. represented by different colours) and on the *P*-Value range obtained over the golden standard calls (i.e. horizontal axis).

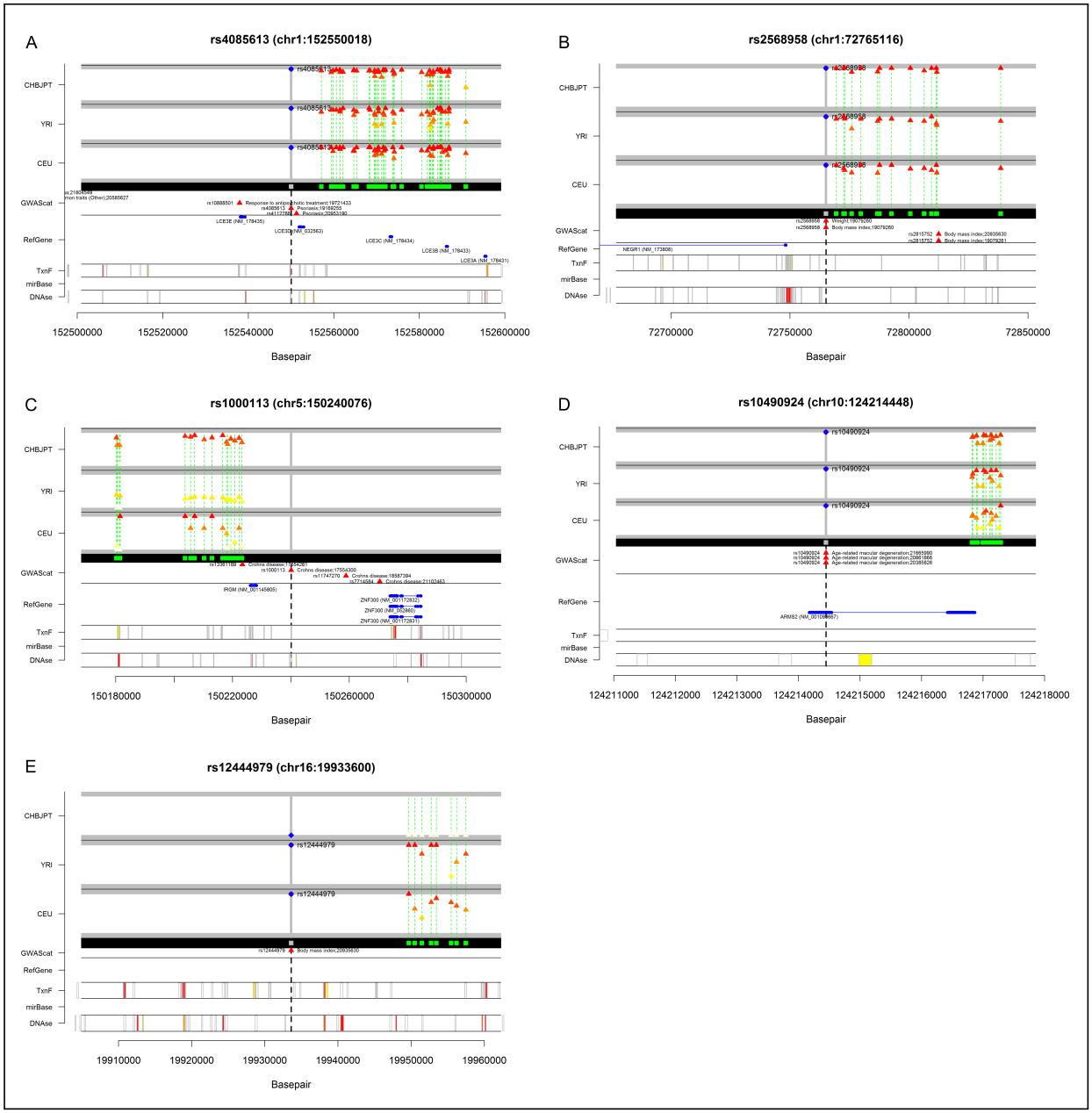


Figure A.9: *Previously reported CNV associations detected by LD analysis between GStream CNV genotypes and trait-associated SNPs.* (A) *LCE* gene cluster deletion associated with Psoriasis risk. (B) *NEGR1* deletion associated with body mass index. (C) *IRGM* deletion associated with Chrohn's disease. (D) *ARMS2* deletion associated with age-related macular degeneration. (E) *GPRC5B* upstream deletion associated with body mass index.

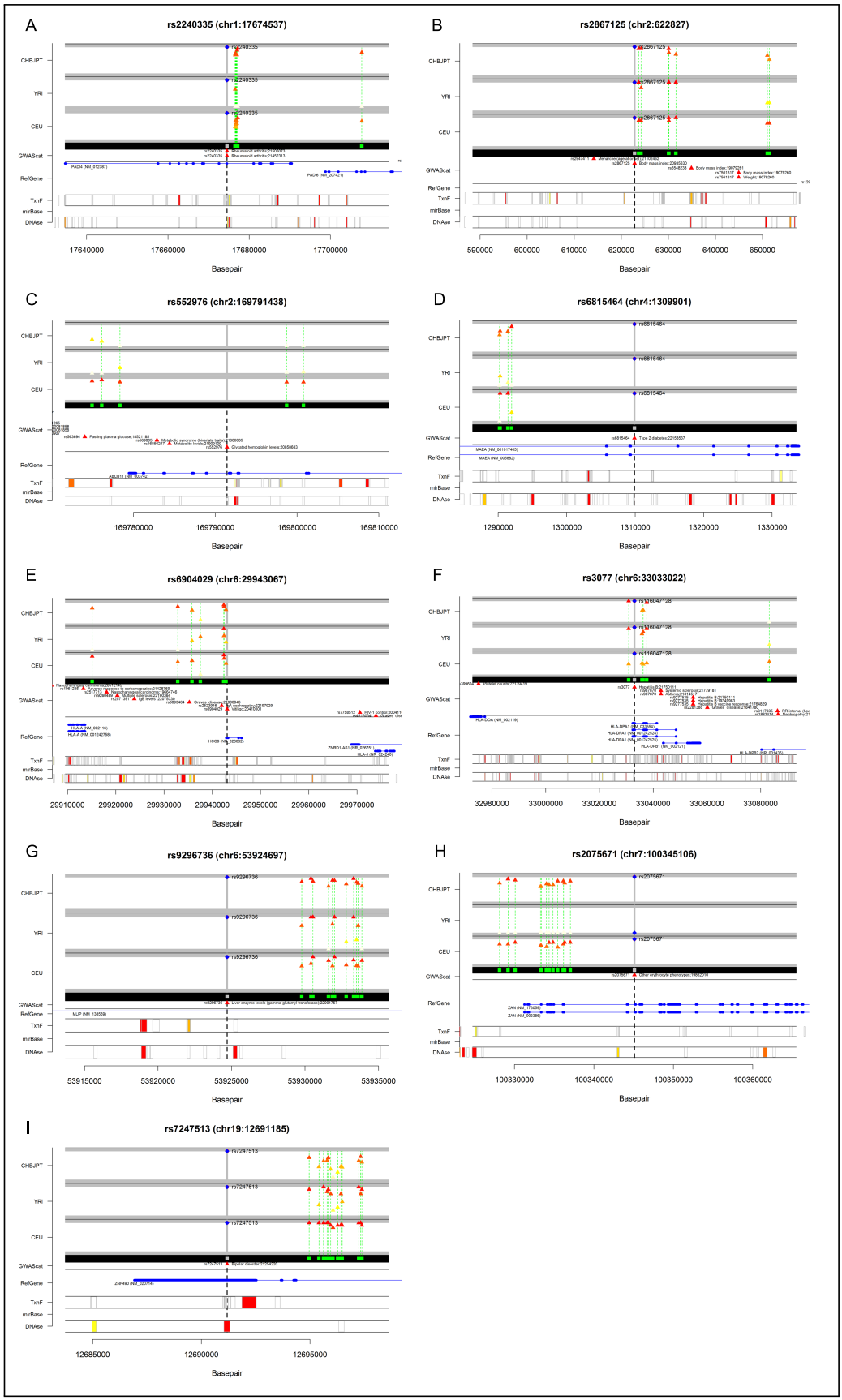


Figure A.10: *Interesting CNV associations detected by LD analysis between GStream CNV genotypes and trait-associated SNPs.* (A) *PADI4* gene deletion associated with Rheumatoid Arthritis. (B) *TMEM18* downstream deletion associated with body mass index. (C) 3'-deletion of gene *ABCB11* associated with glycated hemoglobin levels. (D) *MAEA* gene intron deletion associated with type 2 diabetes. (E) *HCG9* deletion associated with Vitiligo. (F) *HLA-DPA1* deletion associated with Hepatitis B. (G) *MLIP* intron deletion associated with liver enzyme levels. (H) *ZAN* gene deletion associated with red blood cell count. (I) *ZNF490* intron deletion associated with bipolar disorder.

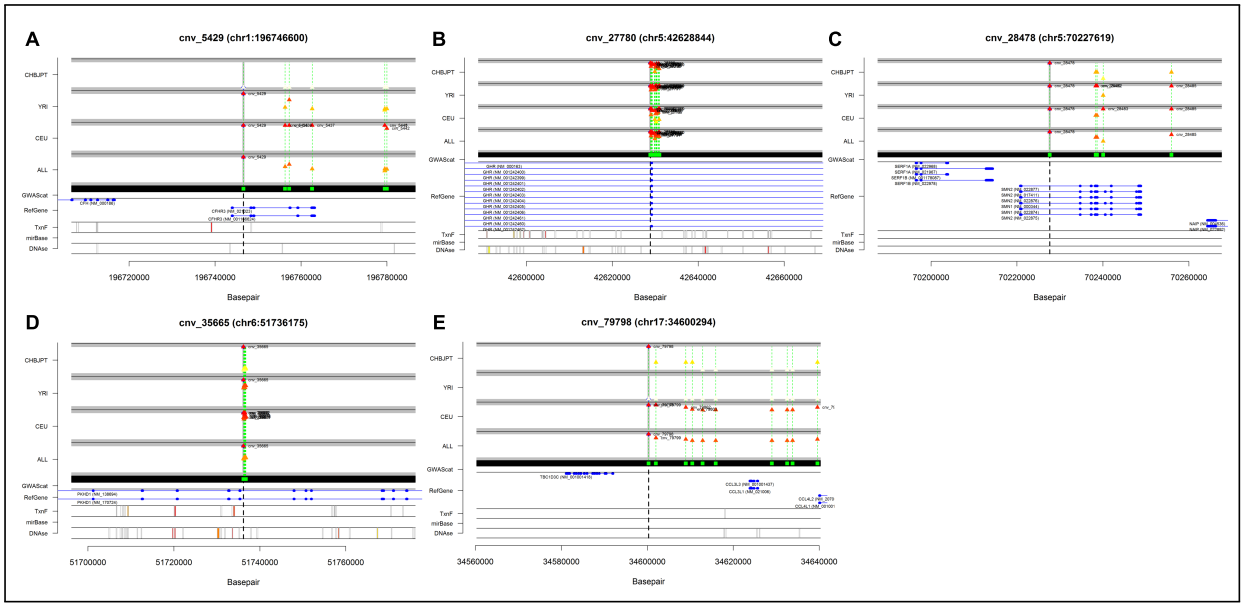


Figure A.11: *GStream detected CNP loci spanning disease-related genes (OMIM) where CNVs have been previously associated with disease.* (A) CNP spanning *CFHR1* and *CFHR3* previously associated to age-related macular degeneration. (B) Deletion of *GHR* exon 3 that has been previously associated with increased responsiveness to growth hormone and Laron dwarfism. (C) Detected *SMN* gene deletion previously associated with spinal muscular atrophy. (D) *PKHD1* deletion associated with polycystic kidney. (E) *CCL3L1/CCL3L3* deletion previously associated with susceptibility to HIV/AIDS.

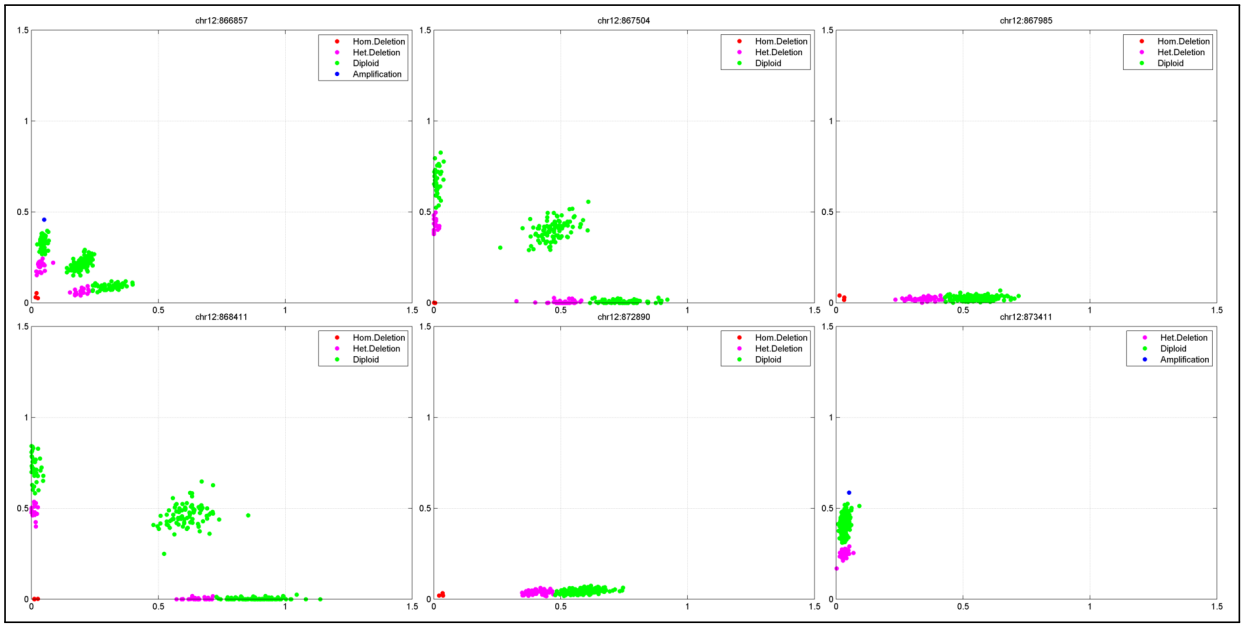


Figure A.12: *GStream calls across consecutive markers spanning the same CNV loci.* These 6 microarray probes cover the same CNV loci but show very different CNV intensity patterns. GStream is completely adapted to these types of variations and its calling procedure is able to obtain very concordant calls when analyzing probes spanning the same CNV.

A.2 Supplementary tables

Online resource
doi:10.1371/journal.pone.0068822.s013
Download link

Table A.1: CNV markers in high LD with trait-associated SNPs reported in the GWAS catalog.

Online resource
doi:10.1371/journal.pone.0068822.s014
Download link

Table A.2: Set of 149 CNV consistent loci spanning OMIM genes.

A.3 GStream algorithm

A.3.1 Input data

The input data required by GStream is provided by a single file and corresponds to channel A and B intensities for each sample at each probe. Each line corresponds to one probe and columns one to three are reserved for the probe annotation data (name, chromosome and basepair). The following columns must contain channel A and B intensities for each sample. Therefore the expected number of columns is $N_{col} = 3 + 2N_S$, where N_S is the number of analyzed samples. We recall that channel A and B intensities are proportional to the number of copies of A and B alleles.

The input file can be easily generated from the Illumina GenomeStudio software by selecting the data fields required by GStream. Figure A.13 shows the appropriate fields that have to be used to export GenomeStudio data to GStream.

Since each probe is processed independently, the details on the algorithm will be specified for only one probe. From here on, channels A and B will be denoted as channel X and Y to avoid confusion between SNP genotypes (AA, AB and BB) and channel names. The following nomenclature will be used to refer to the input data:

- X_n: Channel X intensity for sample n at the processed SNP probe
- Y_n: Channel Y intensity for sample n at the processed SNP probe

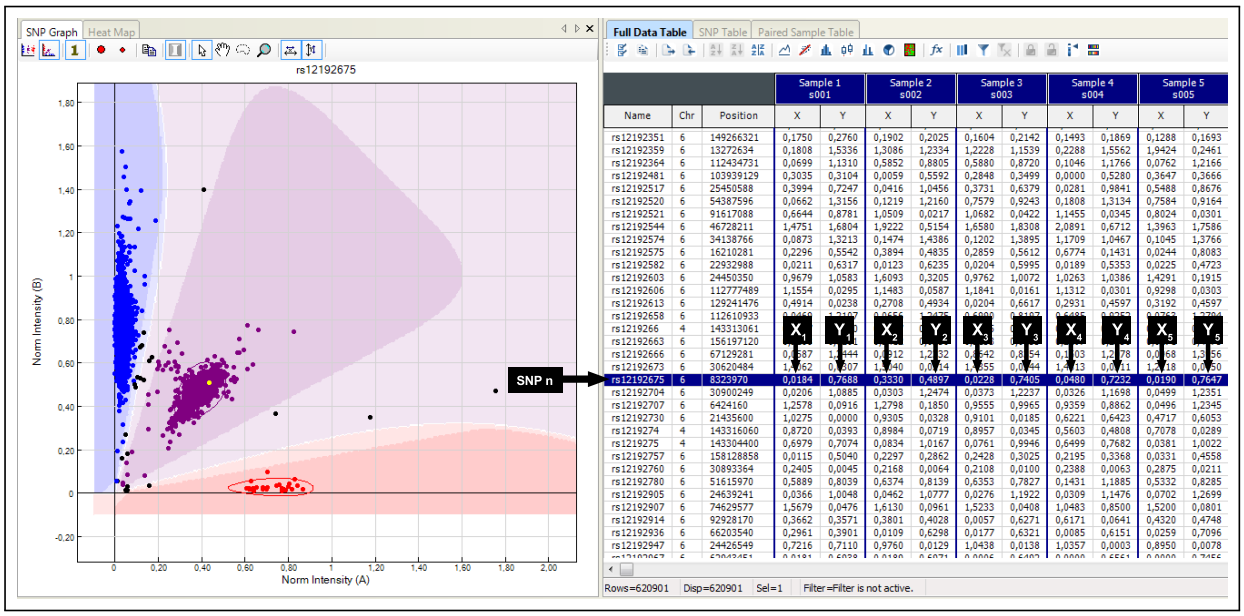


Figure A.13: *GenomeStudio screenshot.* GenomeStudio data table showing the required columns for GStream analysis.

The following sections explain in detail the algorithmic procedures integrated in GStream. These sections follow the processing workflow of GStream (Figure A.14) and are subsequently and independently applied to each SNP probe:

- *Normalization:* This step aims to normalize channel X and Y intensity distributions in order to redress the sensibility bias between both channels.
- *SNP Genotyping:* This step has the objective of assigning a SNP genotype GT_n to each sample n. Samples are clustered in three main groups: AA homozygotes, AB heterozygotes and BB homozygotes ($GT_n \in AA, AB, BB$).
- *CNV Genotyping:* This step assigns a copy number score SC_n to each sample n.

A.3.2 Intensity normalization

This step performs channel intensity normalization in order to equalize both channel intensity distributions. Once this normalization step is performed, a coordinate transformation is applied to obtain the absolute intensities I_n and the allelic frequencies BAF_n. Channel intensity normalization is crucial since the sensibility differences of each SNP probe and channel can lead to bias affecting the genotyping performance.

First, in order to avoid numerical overflows raw channel intensities under a sensitivity threshold (T_S) are set to the minimum allowed value T_S:

$$X'_n = \begin{cases} T_S, & \text{if } X_n < T_S; \\ X_n, & \text{else.} \end{cases} \quad Y'_n = \begin{cases} T_S, & \text{if } Y_n < T_S; \\ Y_n, & \text{else.} \end{cases} \tag{A.1}$$

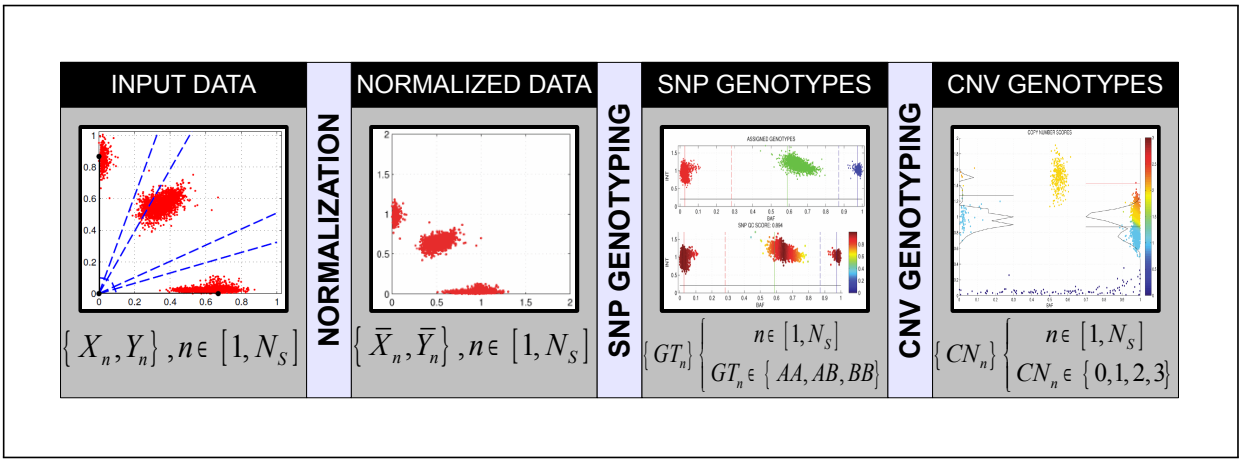


Figure A.14: *GStream workflow.* This schema shows the three main stages of GStream method.

The following step consists of the identification the two candidate homozygote groups, containing the samples that have a high probability of being diploid homozygotes. The candidates must fulfill two conditions:

- Having high absolute intensities ($I_n = X'_n + Y'_n$).
- Having angular coordinates ($BAF_n = \frac{2}{\pi} \arctan \frac{Y'_n}{X'_n}$) next to 0 (AA homozygotes) or 1 (BB homozygotes).

$$n \in AA_{cand} \Leftrightarrow \begin{cases} I_n = X'_n + Y'_n > I_{min} \\ BAF_n < \theta_L \end{cases} \qquad n \in BB_{cand} \Leftrightarrow \begin{cases} I_n = X'_n + Y'_n > I_{min} \\ BAF_n > 1 - \theta_L \end{cases} \tag{A.2}$$

Once candidate samples are detected (Figure A.16A), the algorithm applies different procedures to compute the scaling factors depending on the candidate groups identified:

- AA and BB candidates found If both candidate groups have been detected, the algorithm independently computes the scaling factors α_X and α_Y of each channel. These scaling factors correspond to the maximums of their respective candidate intensity-weighted histograms H_W (Figure A.15) which are computed as follows:

$$\begin{gathered} S(m') = \frac{m'}{10} \max_{n \in AA_{cand}} (I_n) \ , \ m' \in [0 \dots 10] \\ W(m) = 0.5 * (S(m) + S(m+1)), m \in [0 \dots 9] \\ H_W(m) = W(m)^3 * \sum_{n=1}^{N_S} F_I(n \in AA_{cand}) * F_I(S(m) < I_n \leq S(m+1)) \\ \alpha_X = W(\underset{m}{argmax}(H_W(m))) \end{gathered} \tag{A.3}$$

where $F_I(expr) = 1 \leftrightarrow expr = True$,else $F_I(expr) = 0$. α_Y is computed in the same way but changing $AA_{cand} \rightarrow BB_{cand}$. $S(m)$ is a vector containing the centers of the intervals used to compute the histogram and $W(m)$ the weighting factor. α_Y is computed in the same way but changing $AA_{cand} \rightarrow BB_{cand}$ within eq.A.3.

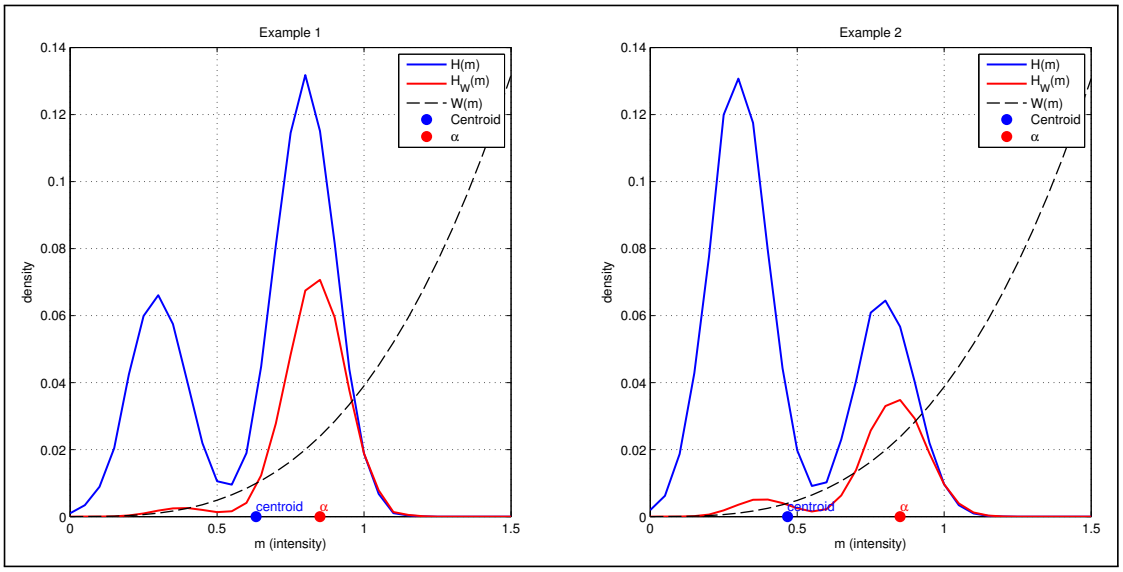


Figure A.15: *Weighted intensity histogram.*. In this figure two candidate homozygote intensity distributions are shown with a typical CNV pattern. The most frequent component corresponds to the two-copy component (higher intensity, left side), while the right side corresponds to the one-copy component. The weighted intensity histogram approach obtains scaling factors corresponding to the average intensity of the two-copy samples.

- Only one candidate group found If only the candidate AA homozygote group is detected (resp. BB), a peak detection algorithm is applied over the BAF histogram. If only one peak is detected, the algorithm assumes that the probe does not capture allelic variation and only detects AA homozygote samples (resp. BB). In that case, both channels are normalized using the corresponding AA homozygote scaling factor α_X (resp. α_Y) as computed in the previous section. If more than one peak is detected the algorithm assumes that peaks correspond to AA homozygote and AB heterozygote clusters (resp. BB and AB). α_X (resp. α_Y) is computed as in the previous section and $\alpha_Y = \alpha_X R_{xy}$ (resp. $\alpha_X = \alpha_Y R_{yx}$) where R_{xy} corresponds to a cross-sensibility ratio computed from the heterozygote sample intensities $R_{xy} = \frac{\overline{Y}_{het}}{\overline{X}_{het}}$ ($R_{yx} = R_{xy}^{-1}$).

- Neither AA nor BB candidates found
  This case is uncommon. α_X and α_Y are computed as follows:

$$\alpha_X = \alpha_Y = \frac{1}{N} \sum_{n=1}^{N_S} I_n \tag{A.4}$$

Once scaling factors α_X and α_Y are found, channel intensities are scaled and new BAF and I coordinates are computed (Figure A.16B):

$$\begin{array}{ll} X'_n = X'_n/\alpha_X & I_n = X'_n + Y'_n \\ & \rightarrow \\ Y'_n = Y'_n/\alpha_Y & BAF_n = \frac{2}{\pi}\arctan\frac{Y'_n}{X'_n} \end{array} \tag{A.5}$$

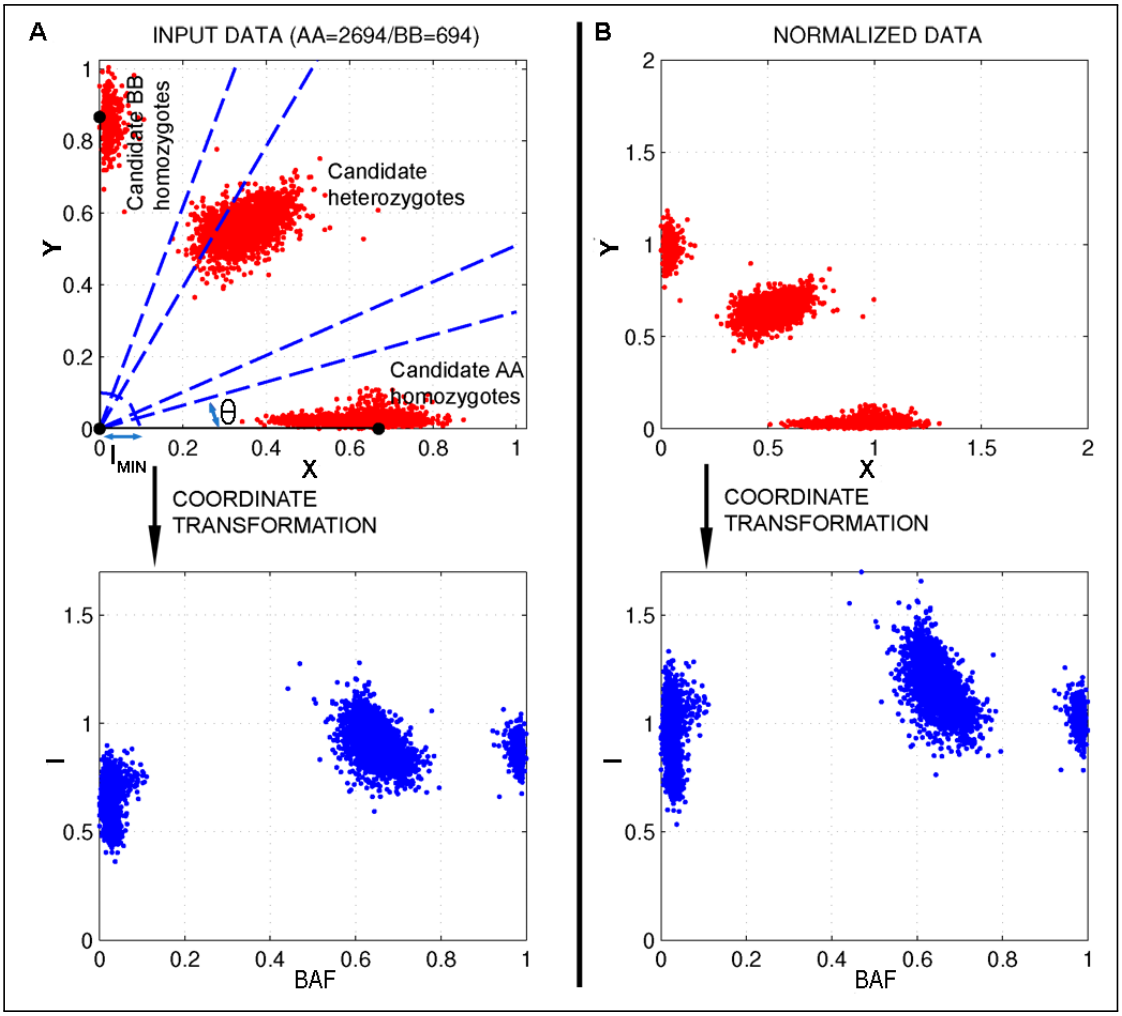


Figure A.16: *Normalization.* (A) This figure shows an example on how homozygote candidates are detected depending on the I_{min} and θ parameters. (B) Shows the resulting intensities after normalization. The intensity differences between AA and BB homozygotes have been corrected and normalized to one.

A.3.3 SNP genotyping algorithm

Once intensities have been normalized, GStream identifies the clusters corresponding to each SNP genotype (AA, AB and BB). In this stage, specific CNV probes from Illumina microarrays are processed as the conventional SNP probes, with the singularity of commonly having only one genotype cluster. The genotyping algorithm is divided into the following steps:

- Zero detection: Absolute intensities (I_n) are used to detect homozygous deletion samples (0 alleles) characterized by low intensities at both channels (Figure A.17). To do that the algorithm sorts the intensities in ascending order and computes the averaged derivative of the resulting vector:

$$j = f(n) \,|I_j \leq I_{j+1}\,, \quad \forall n \in [1 \ldots N_S]$$
$$D(j) = I_{j+1} - I_j \tag{A.6}$$
$$\overline{D}(j) = \frac{D(j-1)+D(j)+D(j+1)}{3}$$

Only sample intensities below the samples intensity average are taken into this calculation. Once $\overline{D}(j)$ has been computed, the second stage consists of identifying its maximum and verifying if it exceeds a predetermined threshold (T_{min}). If true, the zero threshold T_0 is fixed to its corresponding intensity value:

$$T_0 = \begin{cases} 0.5 * (I_{argmax(\overline{D}(j))} + I_{1+argmax(\overline{D}(j))}, & \text{if } max(\overline{D}(j)) > T_{min}; \\ 0, & \text{else.} \end{cases} \tag{A.7}$$

All the samples with an intensity below T_0 are excluded from the following analyses and genotyped as homozygous deletions ($GT_n = 0$ y $SC_n = 0$).

- Limit detection between genotypes: The allele frequency (BAF_n) probability density function (PDF) is estimated computing its histogram $H_{BAF}(m)$. The histogram resolution N_{res} (number of bins) is adjusted depending on the number of analyzed samples between the range $20 < N_{res} < 40$. Once the PDF has been estimated, the maximum peak corresponding to an homozygote cluster is identified ($m_p(1)$). In order to label a peak to an homozygote cluster it must be located at allelic frequencies next to 0 (AA) or 1 (BB). From this peak, a sliding window with a predefined length ($L = \frac{N_{res}}{2}$) is applied over the PDF until another value within the window exceeds the first window value at a predefined distance d_p. When this condition is reached, the BAF value corresponding tho the PDF minimum within the window is set as a genotype limit L_g. This algorithm is applied iteratively until two limits are fixed ($L_{AA|AB}$ y $L_{BB|AB}$) or all the BAF range ($[0, 1]$) has been covered (Figure A.18A). The following pseudocode

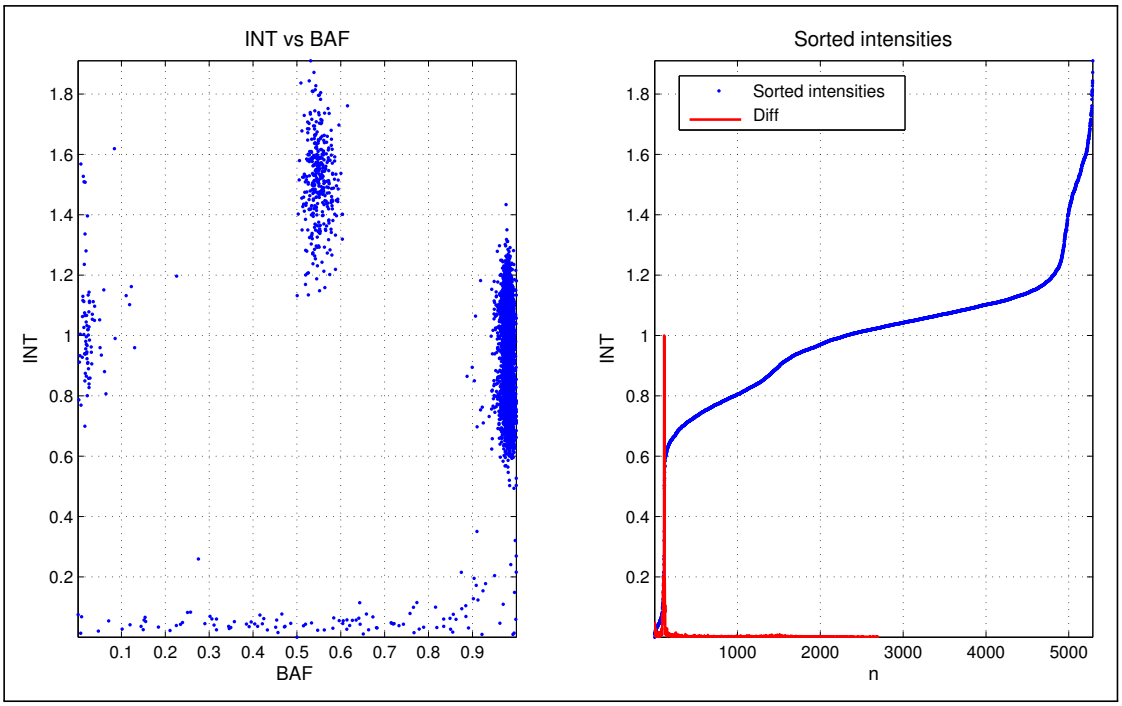


Figure A.17: *Zero detection.* (A) Shows sample BAF and absolute intensity distribution within a probe where homozygote deletions can be observed. (B) Absolute intensities are sorted and peaks are detected over its averaged derivative. If the peak exceeds a threshold, the corresponding intensity will determine the homozygous deletion intensity threshold.

resumes the algorithm:

$H_{BAF}(m),\ m \in [0, 1, ...N_{res}];$ histogram

$m_p(1) = \underset{m}{\text{argmax}}(H_{BAF}(m));$ maximum homozygote peak

$i_{peak} = 2;$

$while(i_{peak} < 3),$ stop condition 1

$for(i = (m_p + 1) : N_{res}),$ sliding window starting point

$m_i = \underset{m \in [i...i+\frac{N_{res}}{2}]}{\text{argmax}} (H_{BAF}(m))$ maximum within the window

$if\,(m_i - m_p(i_{peak} - 1) >= d_p),$ new genotype peak

$m_p(i_{peak}) = m_i;$ new peak position

$i_{peak} = i_{peak} + 1;$

$Lg_{i_{peak}-1} = \underset{m \in [m_p(i_{peak}-1)...m_p(i_{peak})]}{\text{argmin}} (H_{BAF}(m))$ genotype limit

$break;$ search new peak

end

$if\,(i == N_{res}),\ break;\ end$ stop condition 2

end

end

The limits between genotypes determine the BAF ranges assigned to each genotype and each sample will be genotyped accordingly (Figure A.18B):

$$GT_n = \begin{cases} 0, & \text{if } I_n < T_0; \\ 1, & \text{if } I_n \geq T_0 \ \& \ BAF_n < L_{AA|AB}; \\ 2, & \text{if } I_n \geq T_0 \ \& \ L_{AA|AB} \leq BAF_n < L_{BB|AB}; \\ 3, & \text{if } I_n \geq T_0 \ \& \ BAF_n \geq L_{BB|AB}. \end{cases} \tag{A.8}$$

- Re-Genotyping: If the number of detected clusters is less than three, each cluster is reanalyzed with a better resolution (increasing the number of bins used for the PDF estimation) with the purpose of identifying subclusters corresponding to different genotypes. This method avoids common errors seen in other algorithms where, for example, the genotypes corresponding to probes with high discordant sensibilities between channels are incorrectly assigned.

- Scoring: Finally, a probe quality score and an individual sample genotyping score are computed (Figure A.18C). The global score is proportional to the average standard deviation between the BAF values assigned to each genotype, while the individual scores correspond to the distance between BAF sample value and its assigned genotype cluster center normalized by the distance between genotype cluster centers.

A.3.4 CNV genotyping algorithm

GStream uses normalized intensities and SNP genotypes computed in the SNP genotyping stage to identify the presence of deletions and amplifications. These variations are characterized by variable clustering patterns on the intensity probe data (i.e. high frequency CNVs) or by slight deviations from the diploid distribution (i.e. low frequency CNVs).

One of the improvements incorporated by the algorithm is that each SNP genotype cluster is independently analyzed, taking only into account the intensity channel that carries valuable information. This way, the CNV algorithm is divided in four parallel steps (Figure A.19):

- Analysis of channel A intensities from the samples genotyped as AA homozygotes ($X'_{n|GT_n=AA}$).
- Analysis of channel B intensities from the samples genotyped as BB homozygotes ($Y'_{n|GT_n=BB}$).
- Analysis of channel A intensities from the samples genotyped as AB heterozygotes ($X'_{n|GT_n=AB}$).

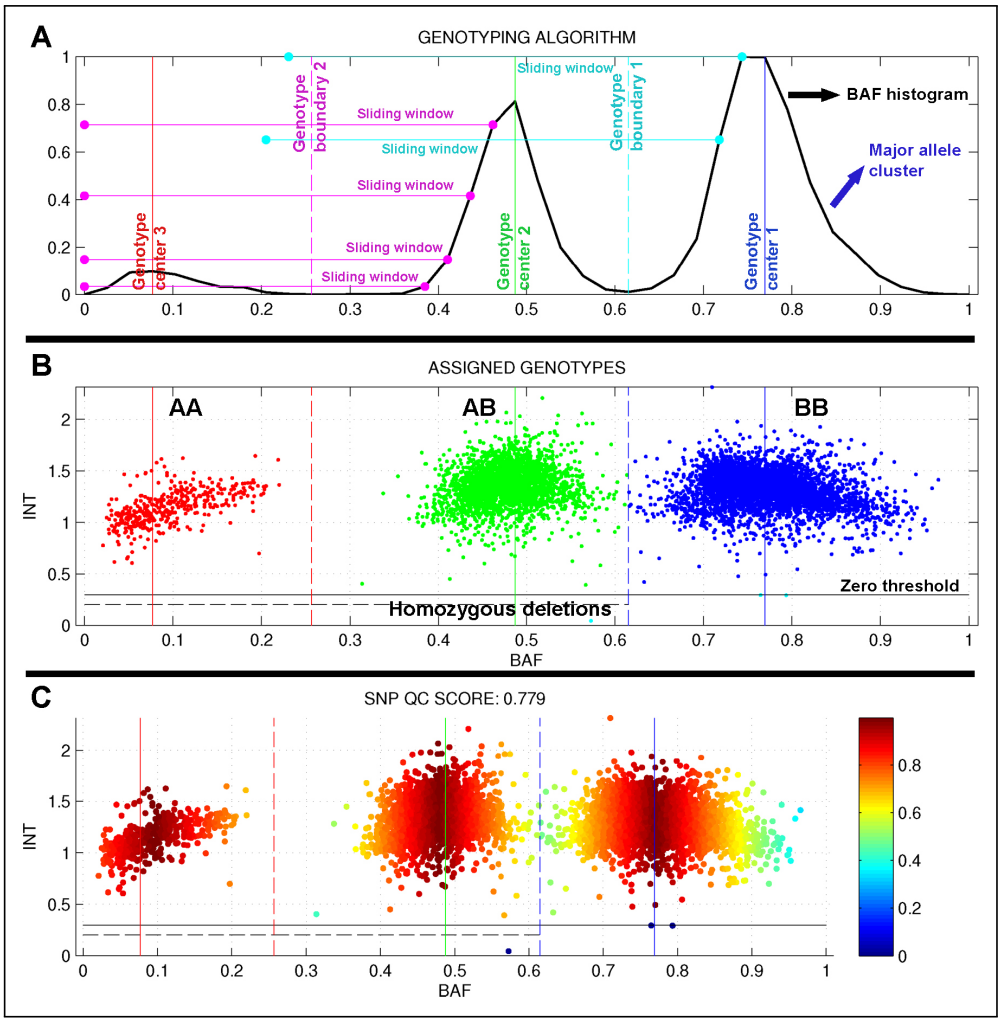


Figure A.18: *Limit detection between genotypes.* (A) Limit detection algorithm over the BAF probability density function. (B) Resulting genotypes assigned by GStream. (C) Sample genotype scores.

- Analysis of channel B intensities from the samples genotyped as AB heterozygotes ($Y'_{n|GT_n=AB}$).

As well as dividing the analysis in four independent steps, the algorithm is based on the following assumptions:

- Homozygous deletions (0 copies) are previously detected during the SNP genotyping stage.
- Due to the technical limitations of genotyping microarrays, the intensity measurements show a saturation effect when amplifications are found. For this reason, intensity clustering patterns corresponding to amplifications are very rare and hard to detect unless they span multiple probes.
- Samples categorized as homozygote samples (i.e. AA and BB) can correspond to heterozygous deletions (i.e. A and B) or amplifications (i.e. AA+ and BB+). Due to the

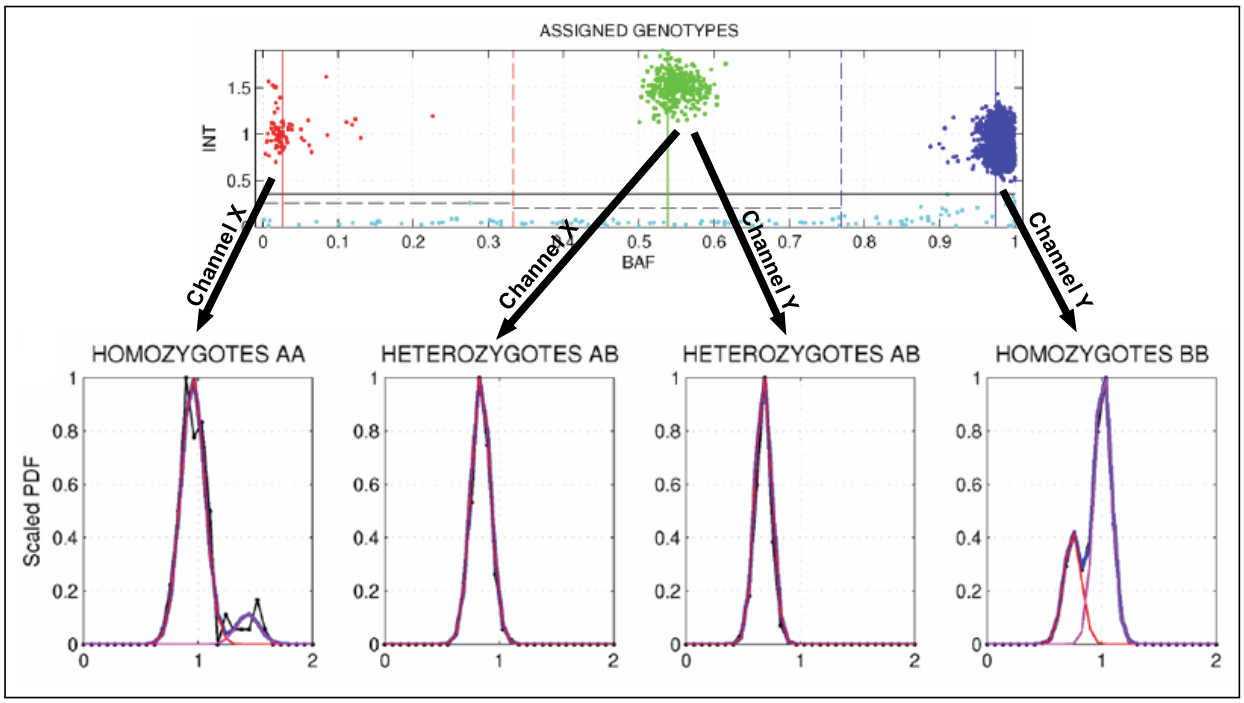


Figure A.19: *Genotype and channel independent analysis.* Each SNP genotype cluster is independently analyzed using the intensity channel that carries the information.

saturation effect the algorithm does not stratify amplifications by the number of allele copies.

- Samples characterized as heterozygotes (i.e. AB) can have two or more copies (i.e. AB, AAB, ABB...). The total number of copies can be inferred by independently computing the number of copies of each allele and then adding the results for each sample.

Below we describe the required steps for determining the CNV genotypes using the SNP genotype data and the normalized channel intensities:

Model selection

The algorithm starts by adjusting two probabilistic models to the channel intensities that carry the allele information corresponding to the SNP genotype cluster that is being analyzed. Due to the mentioned saturation effects, it is very uncommon to observe more than two intensity clusters in microarray data and, for this reason, only two models will be fitted to the intensity data: a one-component $\mathscr{P}_1$ and a two-component $\mathscr{P}_2$ Gaussian mixture model (GMM):

- AA homozygotes $\Longrightarrow \mathscr{P}_1(X_n|n \in AA)$ and $\mathscr{P}_2(X_n|n \in AA)$
- BB homozygotes $\Longrightarrow \mathscr{P}_1(Y_n|n \in BB)$ and $\mathscr{P}_2(Y_n|n \in BB)$

- AB heterozygotes $\Longrightarrow \mathscr{P}_1(X_n|n \in AB)$ and $\mathscr{P}_2(X_n|n \in AB)$
- AB heterozygotes $\Longrightarrow \mathscr{P}_1(Y_n|n \in AB)$ and $\mathscr{P}_2(Y_n|n \in AB)$

These two models will be evaluated over the intensity data to identify which fits better. The procedure is detailed for the channel X over the cluster genotype AA but it can be easily extrapolated to other genotypes and channels ($X_n \rightarrow Y_n$ y $AA \rightarrow AB$ or BB).

The first model is fitted using the mean and the variance of the corresponding intensities:

$$\begin{gathered} \mu_a = \tfrac{1}{N_{AA}} \textstyle\sum_{n \in AA} X'_n \\ \sigma_a^2 = \tfrac{1}{N_{AA}} \textstyle\sum_{n \in AA} (X'_n - \mu_a)^2 \\ X'_{n|n \in AA} \sim \mathscr{P}_1(X) = \mathscr{N}(\mu_a, \sigma_a^2) \end{gathered} \tag{A.9}$$

where N_{AA} refers to the number of samples genotyped as AA. The second model is fitted using the Expectation-Maximization algorithm (EM):

$$X'_{n|n \in AA} \sim \mathscr{P}_2(X) = \omega_1 \mathscr{N}(\mu_{b_1}, \sigma_{b_1}^2) + \omega_2 \mathscr{N}(\mu_{b_2}, \sigma_{b_2}^2) \tag{A.10}$$

The initialization of the 2-component model before applying the EM algorithm is crucial to achieve an optimal convergence of the EM algorithm and to reduce the number of iterations required for convergence. The following initialization method (Figure A.20A) has been developed to ensure these results:

$$\begin{gathered} \mu_b = \begin{cases} [med(X'_{n|n \in AA}) - \frac{\sigma_a}{4} \;\;,\;\; med(X'_{n|n \in AA}) + 2\sigma_a], & \text{if } \sum_{n \in AA}(X'_n - med(X'_n)) > 0; \\ [med(X'_{n|n \in AA}) - 2\sigma_a \;\;,\;\; med(X'_{n|n \in AA}) + \frac{\sigma_a}{4}], & \text{if } \sum_{n \in AA}(X'_n - med(X'_n)) \leq 0; \end{cases} \\ \omega_b = \begin{cases} [\frac{2}{3} \;\;,\;\; \frac{1}{3}], & \text{if } \sum_{n \in AA}(X'_{n_i} - med(X'_{n_i})) > 0; \\ [\frac{1}{3} \;\;,\;\; \frac{2}{3}], & \text{if } \sum_{n \in AA}(X'_{n_i} - med(X'_{n_i})) \leq 0; \end{cases} \\ \sigma_b^2 = [0.1\sigma_a^2 \;\;,\;\; 0.1\sigma_a^2] \end{gathered} \tag{A.11}$$

Once fitted the two models, a set of requirements in order to select the second model have been carefully developed and, only if all of them are accomplished, the two-component model (indicating a pattern corresponding to a common CNV) will be selected. Given the 2-component GMM, its probability density function is defined as follows:

$$\mathscr{P}_2(X) = f_1(X) + f_2(X) = \frac{\omega_1}{\sqrt{2\pi\sigma_{b_1}^2}} \exp\left(\frac{-(x-\mu_{b_1})^2}{2\sigma_{b_1}^2}\right) + \frac{\omega_2}{\sqrt{2\pi\sigma_{b_2}^2}} \exp\left(\frac{-(x-\mu_{b_2})^2}{2\sigma_{b_2}^2}\right) \tag{A.12}$$

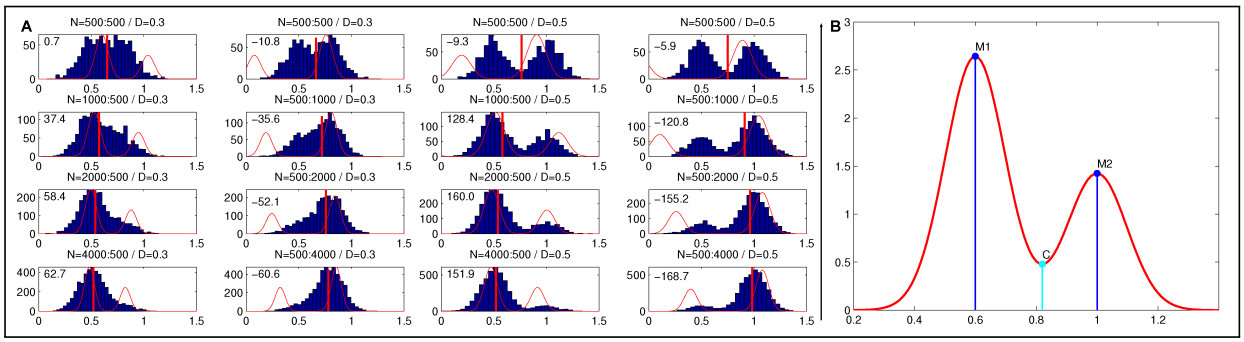


Figure A.20: *Model selection.* (A) Shows the two-component GMM initialization within different intensity distributions. (B) Shows how M_1, M_2 and C are computed.

The X values that maximize each one of the two components (f_1 y f_2) and the X value that minimizes the GMM ($\mathscr{P}_2$) between the two component maximums are defined as follows (Figure A.20B):

$$\begin{aligned} X_{M_1} &= \underset{X}{\operatorname{argmax}}(f_1(X)) \longrightarrow M_1 = f_1(X_{M_1}) \\ X_{M_2} &= \underset{X}{\operatorname{argmax}}(f_2(X)) \longrightarrow M_2 = f_2(X_{M_2}) \\ X_C &= \underset{X \in (X_{M_1}, X_{M_2})}{\operatorname{argmin}} (\mathscr{P}_2(X)) \longrightarrow C = \mathscr{P}_2(X_C) \end{aligned} \tag{A.13}$$

Once these values are computed, the five requirements that must be accomplished for selecting the 2-component model are:

- $\omega_{b_i} > 0.04\ (0.2)\ ,\ i \in [1,2]$
- $f_1(X_{M_2}) < M_1$ y $f_2(X_{M_1}) < M_2$
- $\frac{C}{min(M_1,M_2)} < 0.8\ (0.4)$
- $\frac{X_{M_1}}{X_{M_2}} < 0.85$
- $\overline{D}_{KL} = \max(\frac{1}{2}(\frac{\sigma_{b_1}^2}{\sigma_{b_2}^2} + \frac{(\mu_{b_2}-\mu_{b_1})^2}{\sigma_{b_2}^2} - 1), \frac{1}{2}(\frac{\sigma_{b_2}^2}{\sigma_{b_1}^2} + \frac{(\mu_{b_1}-\mu_{b_2})^2}{\sigma_{b_1}^2} - 1)) > 2\ (6)$

The bracketed values refer to the requirements when analyzing intensities of heterozygote clusters. These values are more restrictive due to their higher variance and to their lower likelihood of having copy number patterns.

Component labeling

Once the models corresponding to each set of intensities and genotypes have been determined, the algorithm must assign a copy number label to each model component. If the

one-component model has been selected, only one label will be required, while the two-component model will require two:

$$\begin{aligned}
&\text{AA Homozygotes} \begin{cases} \mathscr{P}_1(X) \text{ selected} \Longrightarrow CN_{AA}(X) \\ \mathscr{P}_2(X) \text{ selected} \Longrightarrow CN^1_{AA}(X) \text{ and } CN^2_{AA}(X) \end{cases} \\
&\text{BB Homozygotes} \begin{cases} \mathscr{P}_1(Y) \text{ selected} \Longrightarrow CN_{BB}(Y) \\ \mathscr{P}_2(Y) \text{ selected} \Longrightarrow CN^1_{BB}(Y) \text{ and } CN^2_{BB}(Y) \end{cases} \\
&\text{AB Heterozygotes} \begin{cases} \mathscr{P}_1(X) \text{ selected} \Longrightarrow CN_{AB}(X) \\ \mathscr{P}_2(X) \text{ selected} \Longrightarrow CN^1_{AB}(X) \text{ and } CN^2_{AB}(X) \\ \mathscr{P}_1(Y) \text{ selected} \Longrightarrow CN_{AB}(Y) \\ \mathscr{P}_2(Y) \text{ selected} \Longrightarrow CN^1_{AB}(Y) \text{ and } CN^2_{AB}(Y) \end{cases}
\end{aligned} \tag{A.14}$$

This procedure applies different methods depending on the analyzed SNP genotype:

- Homozygotes
  The procedure is detailed for the analysis of channel X intensities over the samples genotyped as AA (equivalent to the analysis of channel Y intensities over the BB samples).
  - $\mathscr{P}_1(X)$ *selected:* Since only one component has been detected, a unique label will be required. The default label assigned to this case is $CN_{AA}(X) = 2$ (the common diploid state) unless a high number of homozygous deletions have been detected. In this case, the assigned label will be $CN_{AA}(X) = 1$ to ensure Hardy-Weinberg equilibrium ($f(--) > f(A-) > f(AA)$).
  - $\mathscr{P}_2(X)$ *selected:* Since two components have been detected, two labels will be required and the weights of each component (ω_1 and ω_2) are expected to be proportional to the frequencies of each group. Labels are assigned to ensure Hardy-Weinberg equilibrium (Figure A.21):

$$[CN^1_{AA}(X)\,,\, CN^2_{AA}(X)] = \begin{cases} [1\ ,\ 2)], & \text{if } \omega_1 < \omega_2\text{; (Figure A.21A)} \\ [1\ ,\ 2], & \text{if } \omega_1 \geq \omega_2 \text{ and } f(--)\uparrow\text{; (Figure A.21B)} \\ [2\ ,\ 3], & \text{if } \omega_1 \geq \omega_2 \text{ and } f(--)\downarrow\text{; (Figure A.21C)} \end{cases} \tag{A.15}$$

- *Heterozygotes*
  When analyzing heterozygotes the common state is AB, which means one copy per intensity channel. Then, when $\mathscr{P}_1(X)$ or $\mathscr{P}_1(Y)$ are selected the default labels are

$CN_X = 1$ or $CN_Y = 1$. When $\mathscr{P}_2(X)$ or $\mathscr{P}_2(Y)$ are selected the default labels are $[CN^1_{AB}(X)\,,\ CN^2_{AB}(X)] = [1,2]$ o $[CN^1_{AB}(Y)\,,\ CN^2_{AB}(Y)] = [1,2]$.

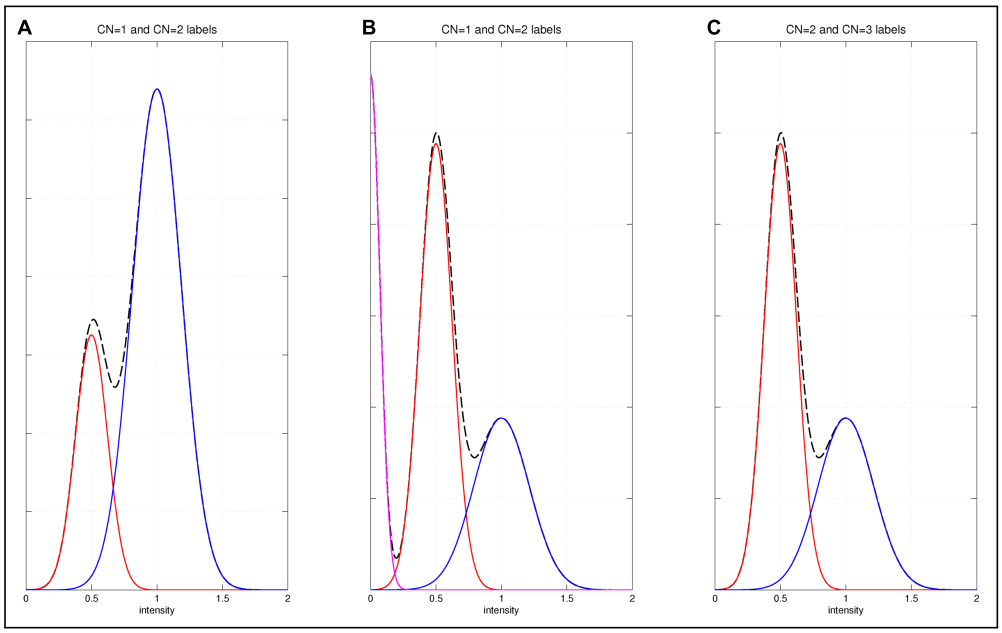


Figure A.21: *Component labeling.* These figures show three examples of how labels are assigned to ensure Hardy-Weinberg equilibrium.

Scoring

Sample scores will be assigned depending on their genotypes and their intensity likelihood relative to each model component. Depending on the SNP genotype scores will be computed as follows (Figure A.22):

- AA Homozygotes:
  - $\mathscr{P}_1(X)$: If the one-component model has been selected, the algorithm assigns to each sample an score proportional to its deviation with respect to the component mean:

$$SC_n = \begin{cases} CN_{AA}(X) + \frac{X_n - \mu_a}{8\sigma_a}, & \text{si } |\frac{X_n - \mu_a}{\sigma_a}| < 8; \\ CN_{AA}(X) + 1, & \text{si } \frac{X_n - \mu_a}{\sigma_a} > 8; \\ CN_{AA}(X) - 1, & \text{si } \frac{X_n - \mu_a}{\sigma_a} < -8; \end{cases} \tag{A.16}$$

  - $\mathscr{P}_2(X)$: If the two-component model has been selected, each sample is scored depending on its intensity and its relative position with respect to the mean of

each component (μ_{b_1} and μ_{b_2}):

$$SC_n = \begin{cases} CN^1_{AA}(X) - 1, & \text{si } X_n < \mu_{b_1} - \frac{8\sigma}{b_1}; \\ CN^1_{AA}(X) + \frac{X_n - \mu_{b_1}}{8\sigma_{b_1}}, & \text{si } \mu_{b_1} - \frac{8\sigma}{b_1} \leq X_n < \mu_{b_1}; \\ \frac{f_1(X_n)}{f_1(X_n)+f_2(X_n)} CN^1_{AA}(X) + \frac{f_2(X_n)}{f_1(X_n)+f_2(X_n)} CN^2_{AA}(X), & \text{si } \mu_{b_1} \leq X_n \leq \mu_{b_2}; \\ CN^2_{AA}(X) + \frac{X_n - \mu_{b_2}}{8\sigma_{b_2}}, & \text{si } \mu_{b_2} \leq X_n < \mu_{b_2} + \frac{8\sigma}{b_2}; \\ CN^2_{AA}(X) + 1, & \text{si } X_n > \mu_{b_2} + \frac{8\sigma}{b_2}; \end{cases} \tag{A.17}$$

- BB Homozygotes:
  Same than previous but replacing: $AA \rightarrow BB$, $X \rightarrow Y$ and $X_n \rightarrow Y_n$.

- AB Heterozygotes:
  Para $W \in [X, Y]$:

  – $\mathscr{P}_1(W)$: If the one-component model has been selected, the algorithm assigns the same score to all the samples corresponding to the component label $CN_{AB}(W)$.

$$SC^W_n = CN_{AB}(W)\,,\ \forall n \in AB \tag{A.18}$$

  – $\mathscr{P}_2(X)$: If the two-component model is selected, each sample is scored depending on its intensity and its relative position with respect to the mean of each component (μ_{b_1} and μ_{b_2}):

$$SC^W_n = \begin{cases} CN^1_{AB}(Z), & \text{si } Z_n < \mu_{b_1}; \\ \frac{f_1(Z_n)}{f_1(Z_n)+f_2(Z_n)} CN^1_{AB}(Z) + \frac{f_2(Z_n)}{f_1(Z_n)+f_2(Z_n)} CN^2_{AB}(Z), & \text{si } \mu_{b_1} \leq Z_n \leq \mu_{b_2}; \\ CN^2_{AB}(Z), & \text{si } \mu_{b_2} \leq Z_n; \end{cases} \tag{A.19}$$

The final score of heterozygote samples will be computed as: $SC_n = SC^X_n + SC^Y_n$

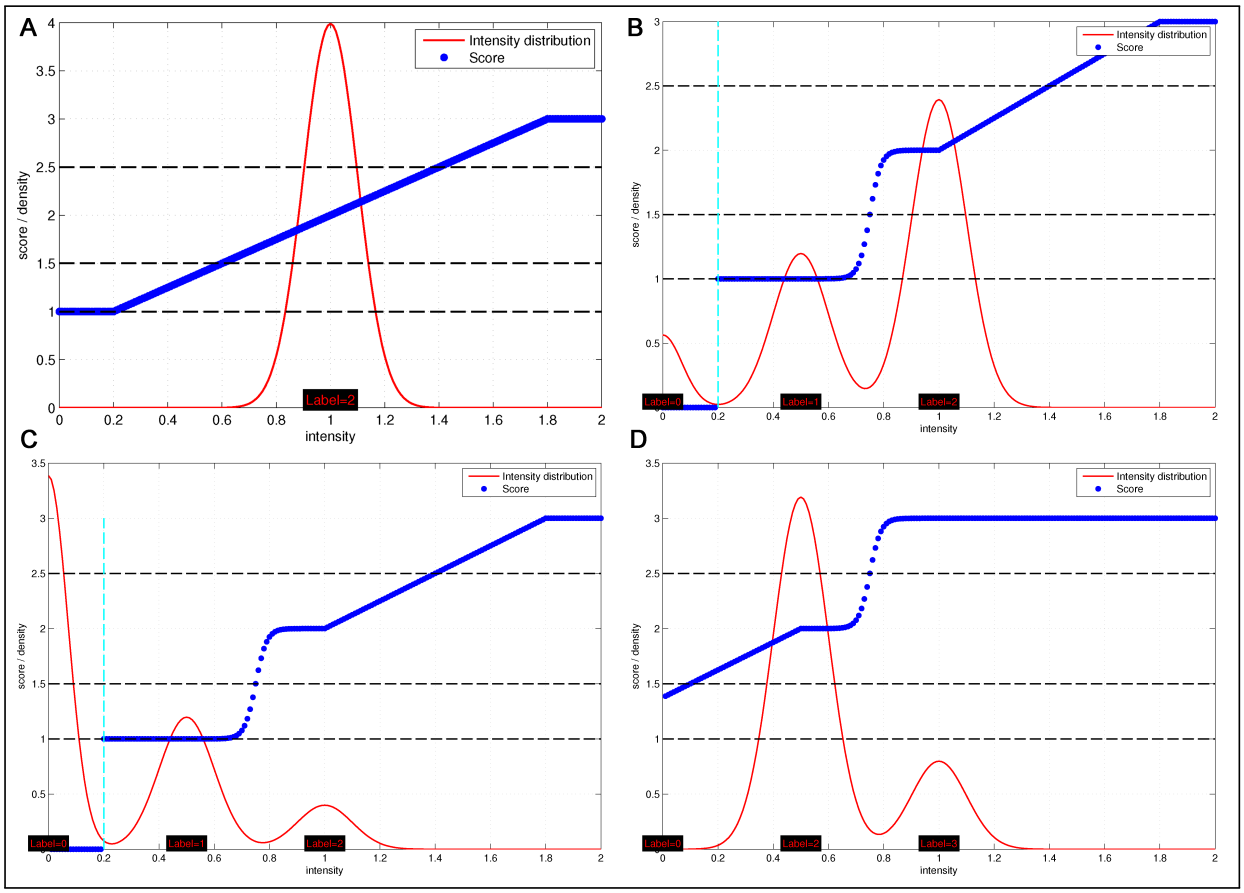


Figure A.22: *Scoring.* These figures show how the scores are assigned depending on the labels and the number of components of the selected model (red line). Blue points show the relationship between sample intensities (horizontal axis) and sample scores (vertical axis). The vertical cyan line represents T_0, the 0-copies intensity threshold computed in the SNP genotyping stage, while the horizontal black-dotted lines represent the limits between discrete copy number assignments.

B | Supplementary Data of FOCUS

B.1 Supplementary figures

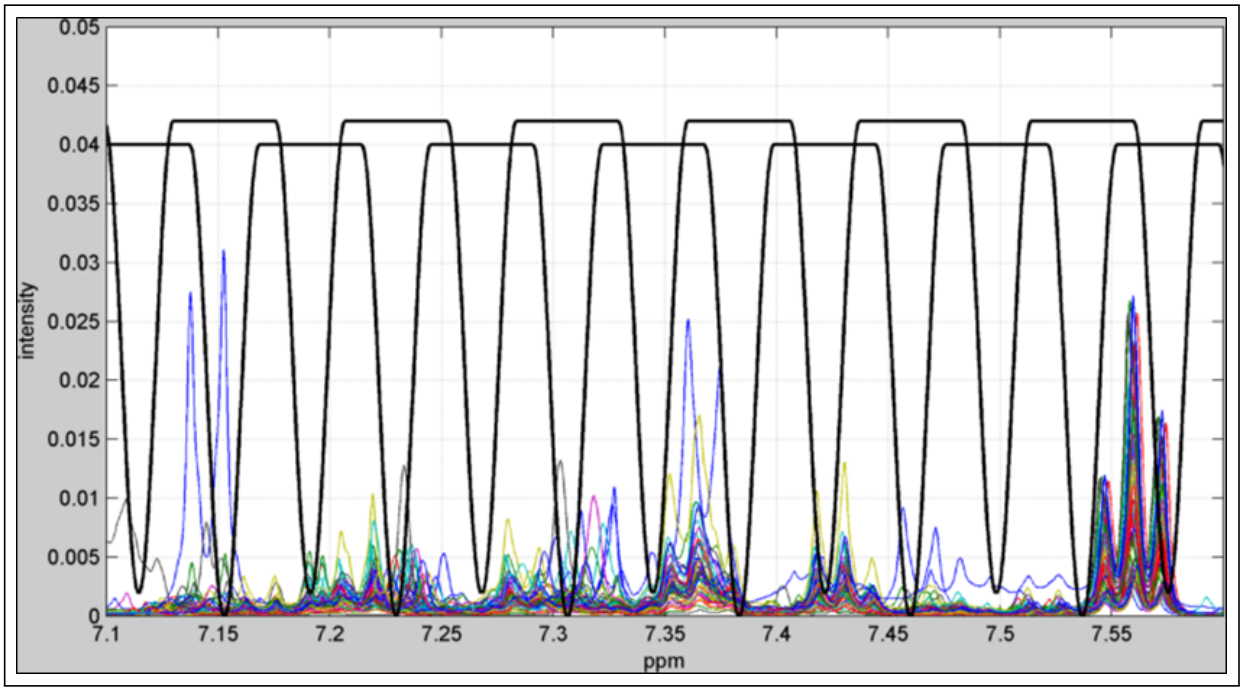


Figure B.1: *Moving window analysis for unsupervised analysis.* This figure shows how a whole dataset of sample spectra is divided in overlapping segments in order to perform an unsupervised analysis. Segment length and overlap can be defined by the user. In this case a segment overlap of 50% and a segment length of 0.08 ppm have been defined.

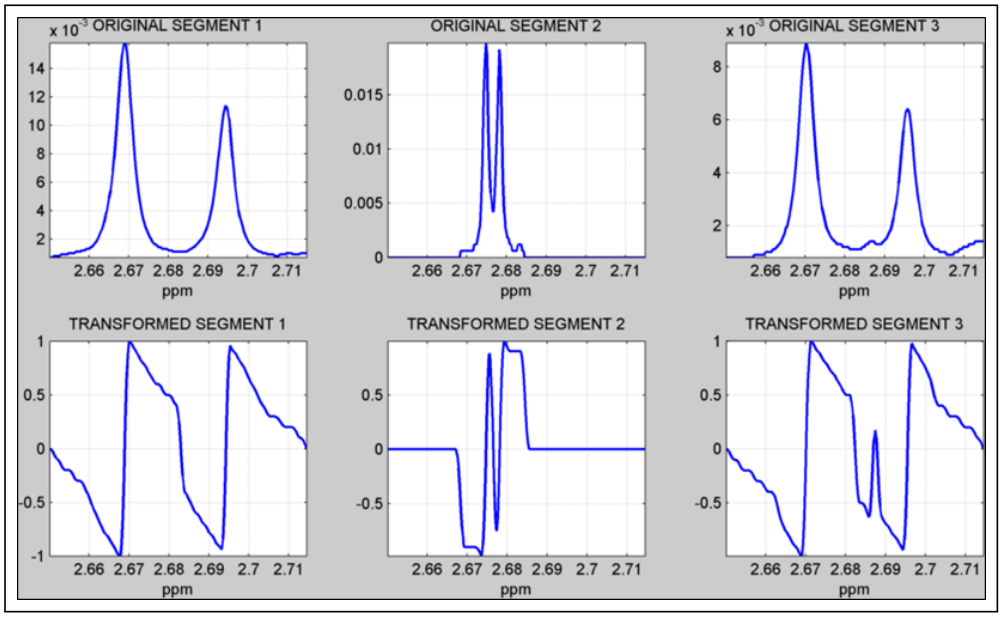


Figure B.2: *Intensity-Weight Signal Transformation.* This figure shows three example spectral segments and their respective transformed segments. As described, this transformation reduces and equalizes the intensity ratios between peaks. It also generates high intensity deltas within the signal peak positions.

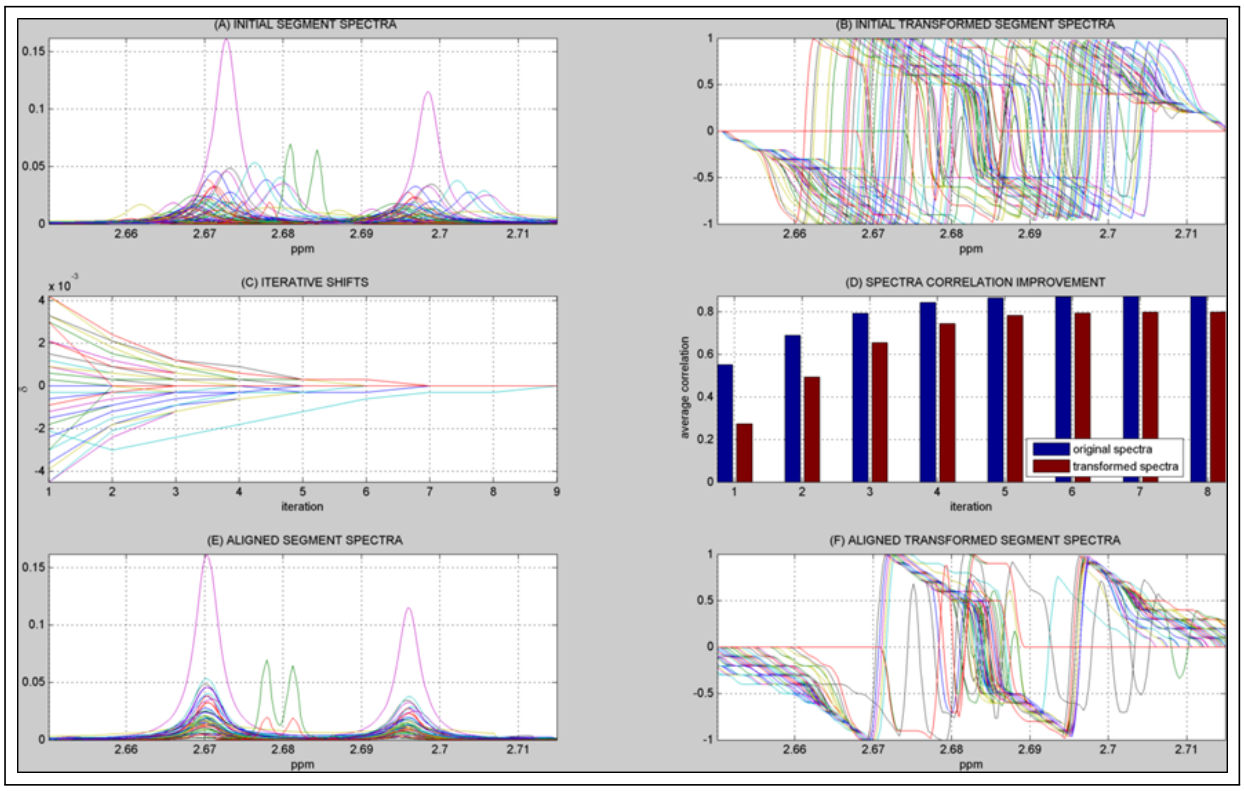


Figure B.3: *Recursive Unreferenced Alignment.* This figure shows a segment alignment example. Figures above represent the input segment spectra (A) and their intensity-weight transforms (B). Figure (C) shows the shift applied to each sample at each iteration and figure (D) the average spectra correlation of the original and transformed spectra across each iteration. Finally, figures (E-F) show the set of original and transformed spectra once they have been aligned. As showed by this alignment result, each group of similar spectra has been independently aligned thanks to the correlation threshold imposed on the iterative shift computation.

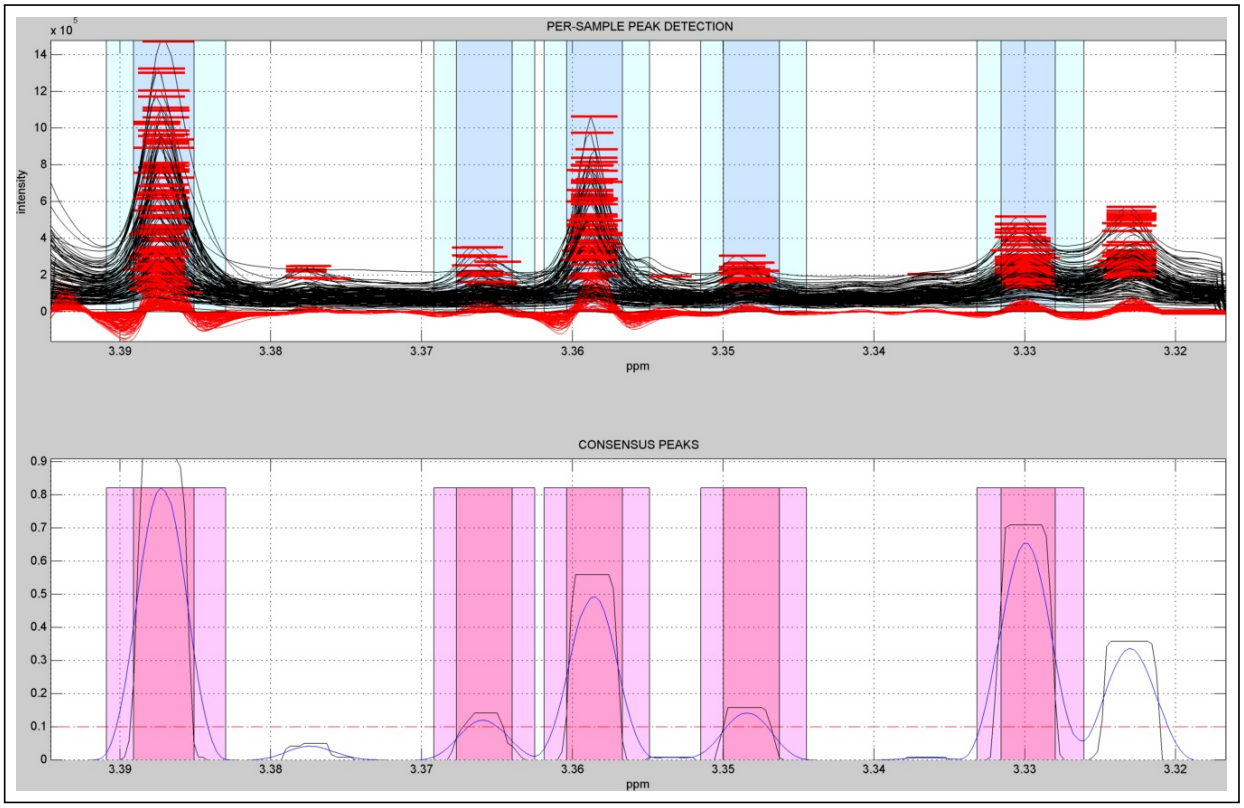


Figure B.4: *FOCUS peak picking.* The upper figure shows the result of applying the FOCUS peak picking algorithm to a spectral segment of the liver extracts NMR dataset. The graphic above shows all the sample spectra (black lines) together with their respective filtered signals (red lines). Horizontal red lines show the peak regions identified on each sample that will be used to build the consensus peak signal showed in the graphic below. Only the peak regions exceeding the frequency threshold (i.e. 0.10) are selected for further analysis (i.e. respectively blue and red shaded regions on the upper and lower plots). The lower figure shows the algorithm adaptation to residual peak unalignment by broadening the integration area due to the variance of the per-sample detected peaks.

FOCUS ANALYSIS RESULTS

STUDY NAME: URINE / 2013-09-18 16:03:03

PEAK SEARCH:

ID_PEAK	INFO	PPM	ID_WINDOW	QC_SC	QC_IvsA	QC_IP	G_95	G_90	G_85	G_80	ID	ID_SCO
p2	±	11.0	3 (11.022-10.944)	0.99	0.44	0.95	G95_17 (1)	G90_21 (1)	G85_24 (1)	G80_22 (1)		
p4	±	9.13	51 (9.144-9.066)	0.73	1.0	0.5	G95_6 (4)	G90_6 (4)	G85_8 (4)	G80_8 (4)	trigonelline-7.0	0.96
p7	±	8.853	58 (8.871-8.793)	0.53	0.97	0.4	G95_6 (4)	G90_6 (4)	G85_8 (4)	G80_8 (4)	trigonelline-7.0	0.95
p6	±	8.843	57 (8.91-8.832)	0.22	0.9	0.25	G95_6 (4)	G90_6 (4)	G85_8 (4)	G80_8 (4)	trigonelline-7.0	0.95
p9	±	8.468	67 (8.519-8.441)	0.89	0.87	0.75	G95_18 (1)	G90_22 (1)	G85_25 (1)	G80_23 (1)	formate-7.0	0.93
p11	±	8.397	69 (8.44-8.362)	0.21	0.91	0.1	G95_19 (1)	G90_23 (1)	G85_26 (1)	G80_24 (1)		
p12	±	8.326	71 (8.362-8.284)	0.48	0.98	0.45	G95_20 (1)	G90_24 (1)	G85_27 (1)	G80_25 (1)		
p13	±	8.273	72 (8.323-8.245)	0.1	0.97	0.1	G95_21 (1)	G90_25 (1)	G85_28 (1)	G80_26 (1)		
p14	±	8.264	73 (8.284-8.206)	-0.43	0.99	0.0	G95_22 (1)	G90_26 (1)	G85_29 (1)	G80_27 (1)		
p15	±	8.248	73 (8.284-8.206)	-0.03	0.98	0.1	G95_23 (1)	G90_27 (1)	G85_30 (1)	G80_28 (1)		
p16	±	8.047	78 (8.088-8.01)	0.12	0.6	0.2	G95_24 (1)	G90_28 (1)	G85_31 (1)	G80_29 (1)		
p17	±	7.963	80 (8.01-7.932)	0.33	0.81	0.25	G95_25 (1)	G90_29 (1)	G85_32 (1)	G80_30 (1)	xanthine-7.0	0.67
p19	±	7.846	83 (7.893-7.815)	0.93	1.0	0.85	G95_2 (10)	G90_2 (10)	G85_2 (11)	G80_3 (11)	hippurate-7.0	1.0
p22	±	7.832	84 (7.854-7.776)	0.93	1.0	0.9	G95_2 (10)	G90_2 (10)	G85_2 (11)	G80_3 (11)	hippurate-7.0	1.0
p23	±	7.757	85 (7.815-7.737)	0.15	0.99	0.1	G95_9 (2)	G90_4 (8)	G85_4 (8)	G80_6 (8)		
p24	±	7.706	86 (7.775-7.697)	0.22	0.75	0.55	G95_26 (1)	G90_30 (1)	G85_33 (1)	G80_31 (1)	pseudouridine-7.0	0.65
p27	±	7.694	88 (7.697-7.619)	0.2	0.75	0.25	G95_27 (1)	G90_31 (1)	G85_34 (1)	G80_32 (1)	pseudouridine-7.0	0.72
p28	±	7.659	88 (7.697-7.619)	0.76	1.0	0.6	G95_2 (10)	G90_2 (10)	G85_2 (11)	G80_3 (11)	hippurate-7.0	1.0
p29	±	7.647	88 (7.697-7.619)	0.83	1.0	0.75	G95_2 (10)	G90_2 (10)	G85_2 (11)	G80_3 (11)	hippurate-7.0	1.0
p33	±	7.634	89 (7.658-7.58)	0.81	0.99	0.75	G95_2 (10)	G90_2 (10)	G85_2 (11)	G80_3 (11)	hippurate-7.0	1.0
p36	±	7.571	91 (7.58-7.502)	0.94	1.0	0.85	G95_2 (10)	G90_2 (10)	G85_2 (11)	G80_3 (11)	hippurate-7.0	0.98
p37	±	7.558	91 (7.58-7.502)	0.96	1.0	0.9	G95_2 (10)	G90_2 (10)	G85_2 (11)	G80_3 (11)	hippurate-7.0	0.98
p38	±	7.546	91 (7.58-7.502)	0.84	1.0	0.8	G95_2 (10)	G90_2 (10)	G85_2 (11)	G80_3 (11)	hippurate-7.0	0.98
p39	±	7.523	92 (7.541-7.463)	0.28	1.0	0.4	G95_2 (10)	G90_2 (10)	G85_2 (11)	G80_3 (11)	hippurate-7.0	0.84
p40	±	7.511	92 (7.541-7.463)	0.03	0.83	0.25	G95_28 (1)	G90_32 (1)	G85_35 (1)	G80_33 (1)	Uracil-7.0	0.68

Figure B.5: *FOCUS summary report.* This figure shows a part of the summary report generated by FOCUS. Each line represents a peak, which is characterized by its position (PPM), the segment window where it was analyzed (ID WINDOW), a set of quality scores (QC-SC,QC-IvsA and QC-IP), a set of group identifiers depending on the grouping correlation threshold (G-95, G-90, G-85 and G-80) and the most probable metabolite identification.

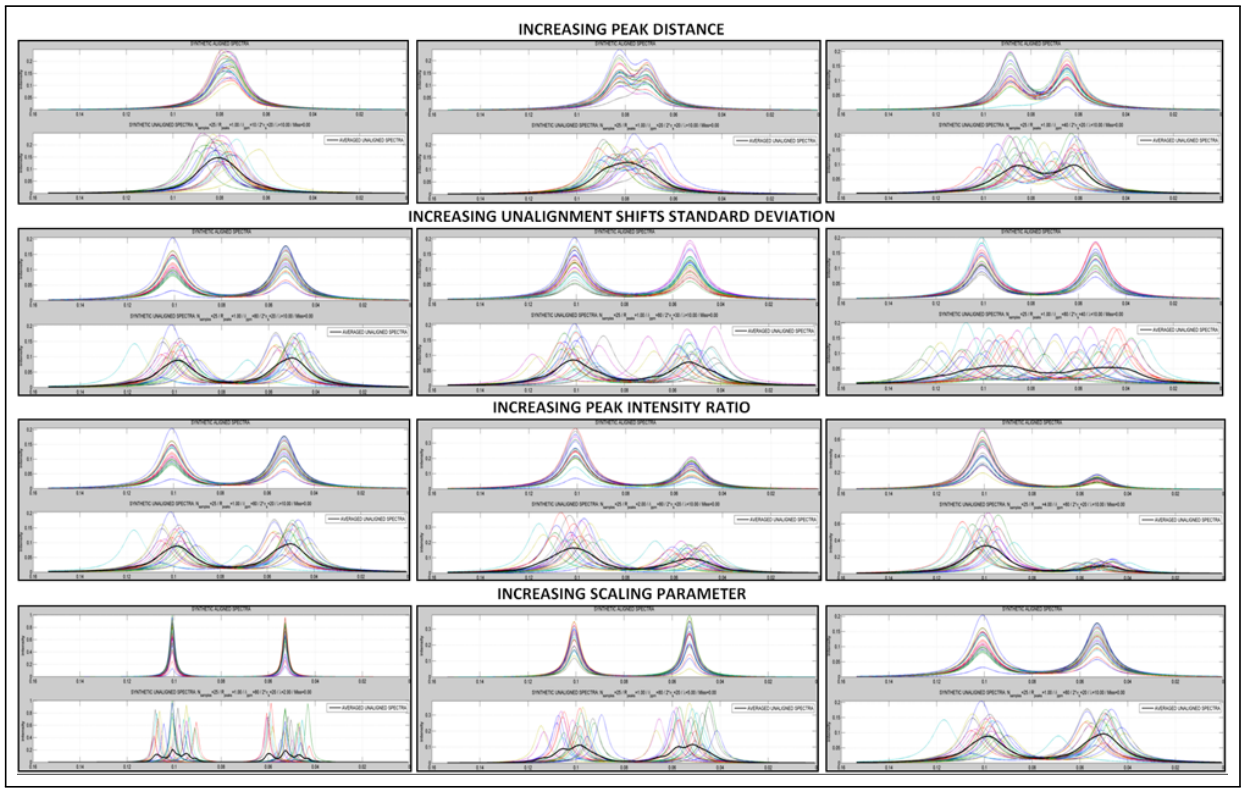


Figure B.6: *Synthetic dataset generation.* This figure shows how the resulting synthetic spectral dataset changes when modifying one parameter while the others are set constants.

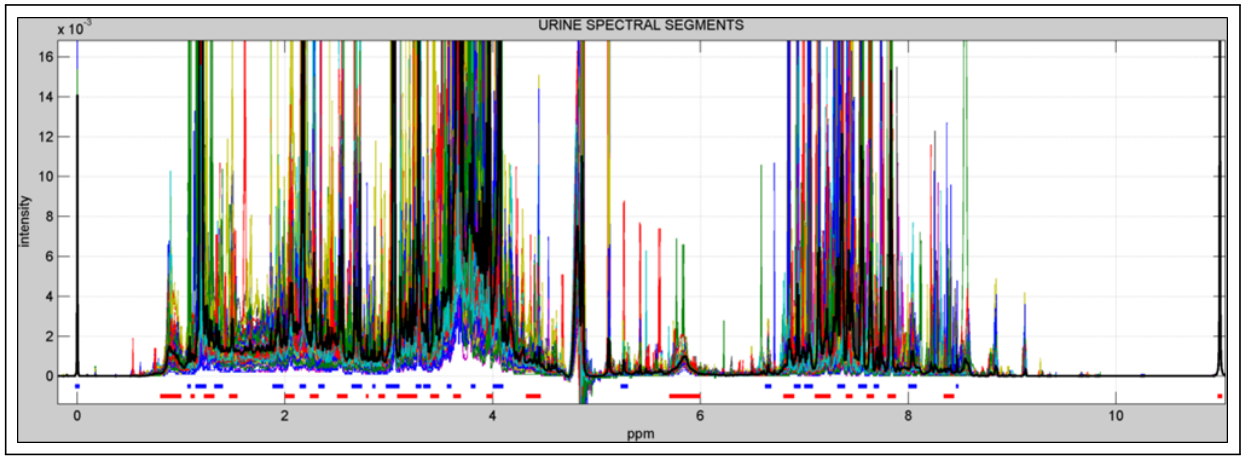


Figure B.7: *Segments for alignment performance evaluation.* This figure shows the 48 manually selected segments to evaluate alignment performance over the dataset composed of 60 human urine spectra.

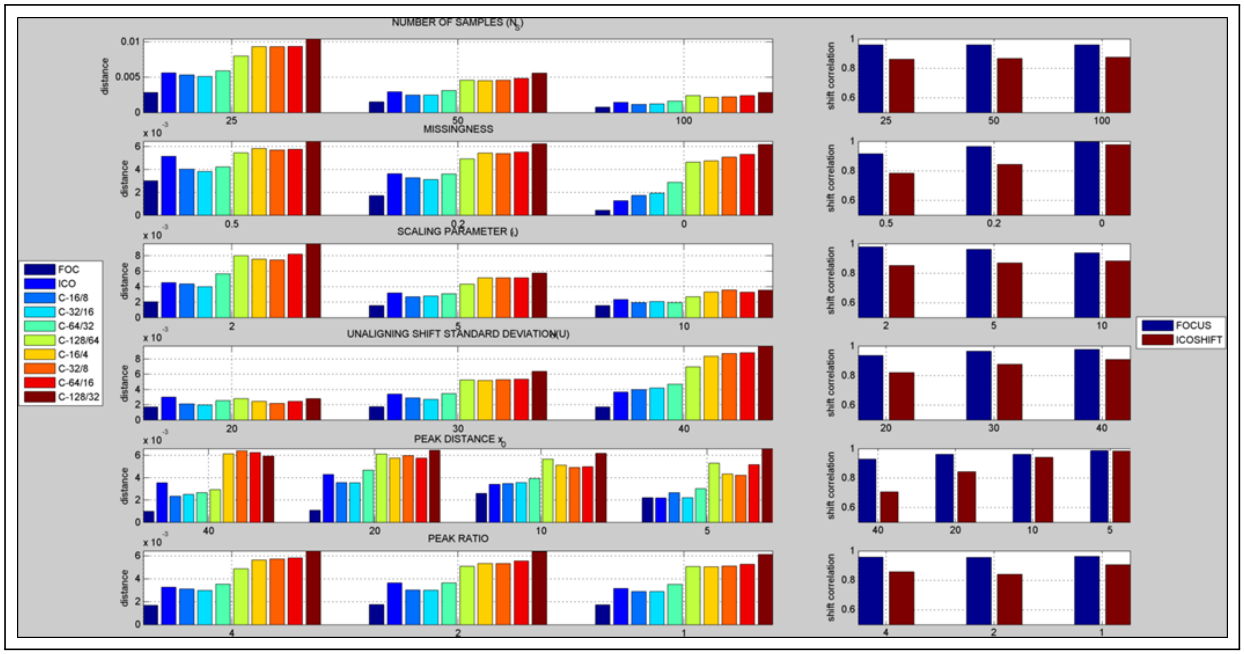


Figure B.8: *Detailed alignment results on the doublet synthetic dataset.* This figure shows the two alignment performance metrics and their dependence on each parameter. Figures at left show correlation matrix distances between the true aligned and the algorithmically aligned spectra using Focus, Icoshift and different COW configurations. Figures at right show correlation coefficients between true and algorithmically computed shifts for Focus and Icoshift.

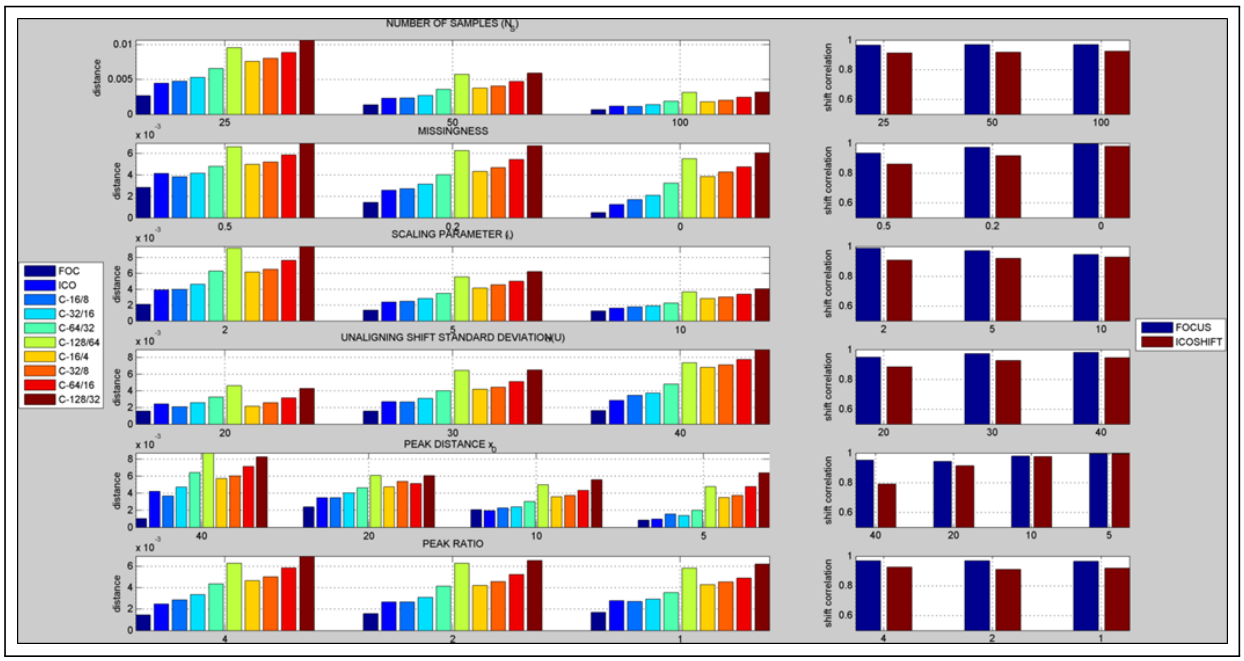


Figure B.9: *Detailed alignment results on the triplet synthetic dataset.* This figure shows the two alignment performance metrics and their dependence on each parameter. Figures at left show correlation matrix distances between the true aligned and the algorithmically aligned spectra using Focus, Icoshift and different COW configurations. Figures at right show correlation coefficients between true and algorithmically computed shifts for Focus and Icoshift.

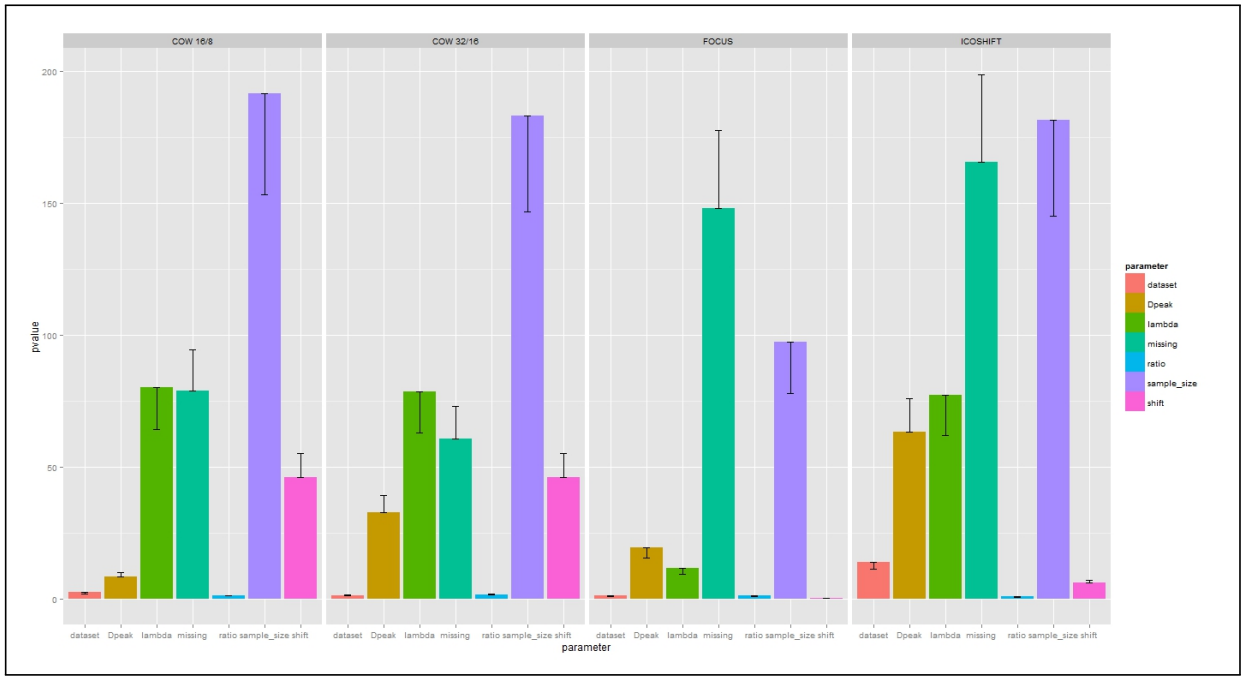


Figure B.10: *Parameter contribution on performance results.* The contribution on alignment performance of each parameter modeling the synthetic spectra is measured for each algorithm. This has been done by using a linear regression test that sets the distance between real and corrected correlation matrices as the dependent variable and all the parameter values as the input variables within all the 972 simulated scenarios. Bars show the $-log_{10}(P-Value)$ of each parameter for each algorithm, black lines above the bars mean positive relation (parameter value increment causes a decrement on the correlation matrix distances -worst alignment-) and black lines under the bars mean negative relation (parameter value increment is related to a decrement on the correlation matrix distances -better alignment-).

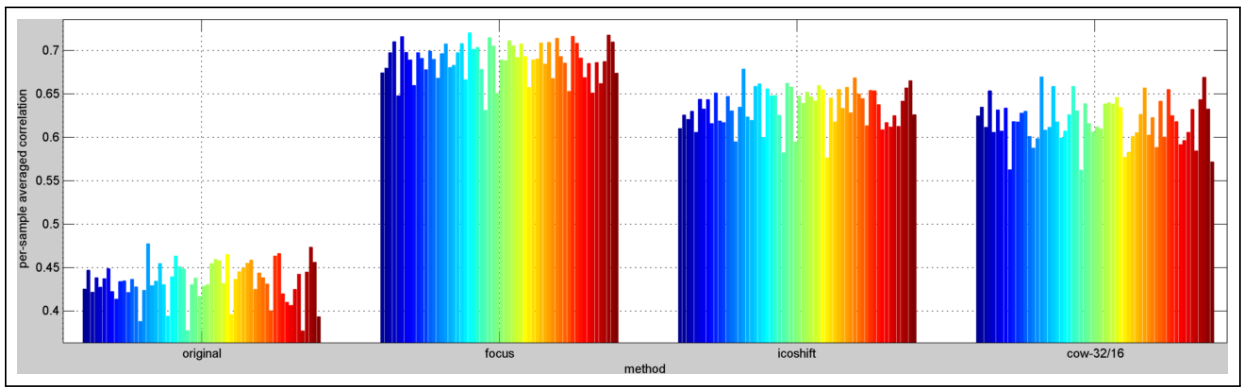


Figure B.11: *Per-sample averaged spectrum correlation.* This figure shows the performance results based on the per-sample averaged spectrum correlation before and after applying the alignment algorithms. FOCUS obtains generalized performance improvements across all the samples.

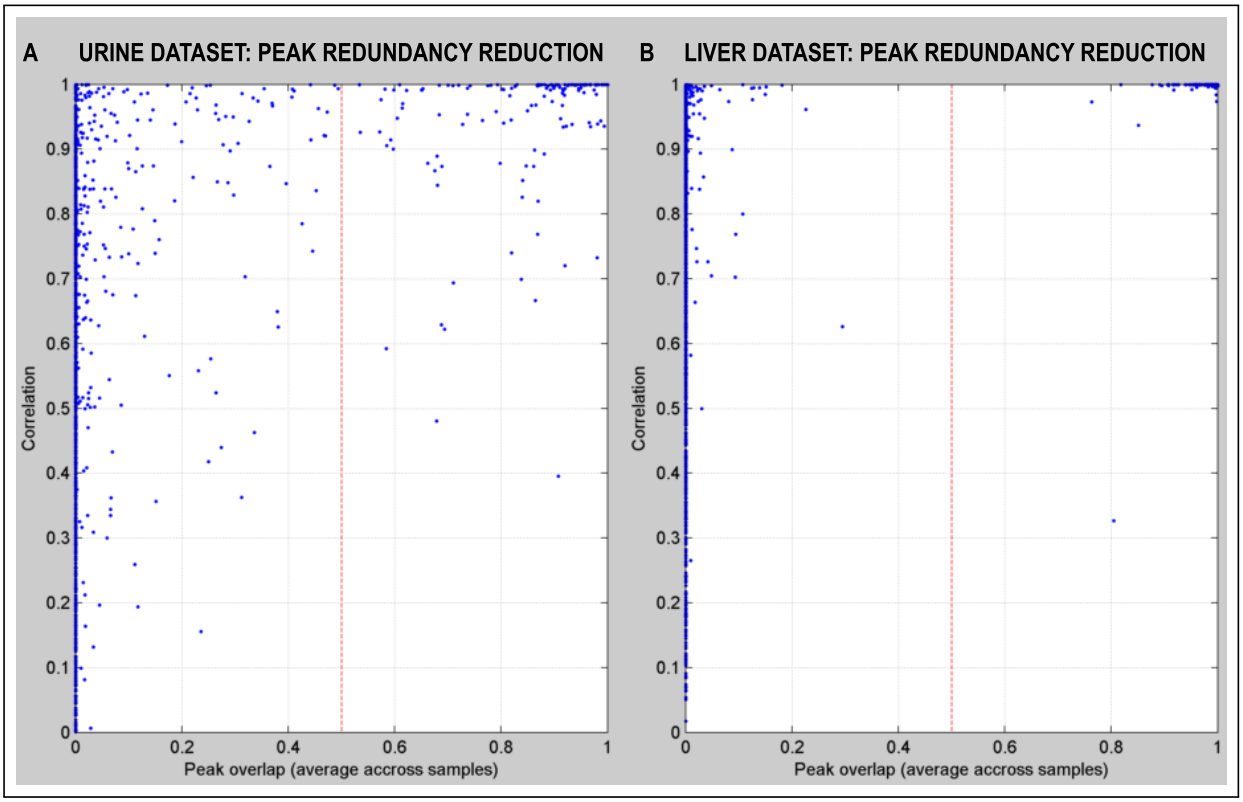


Figure B.12: *Peak redundancy reduction.* This figure shows the peak overlap versus the intensity correlation of the peak pairs of consecutive analysis windows. A peak overlap exceeding 0.5 represents a peak repetition and only one of the two peaks are kept for further analysis.

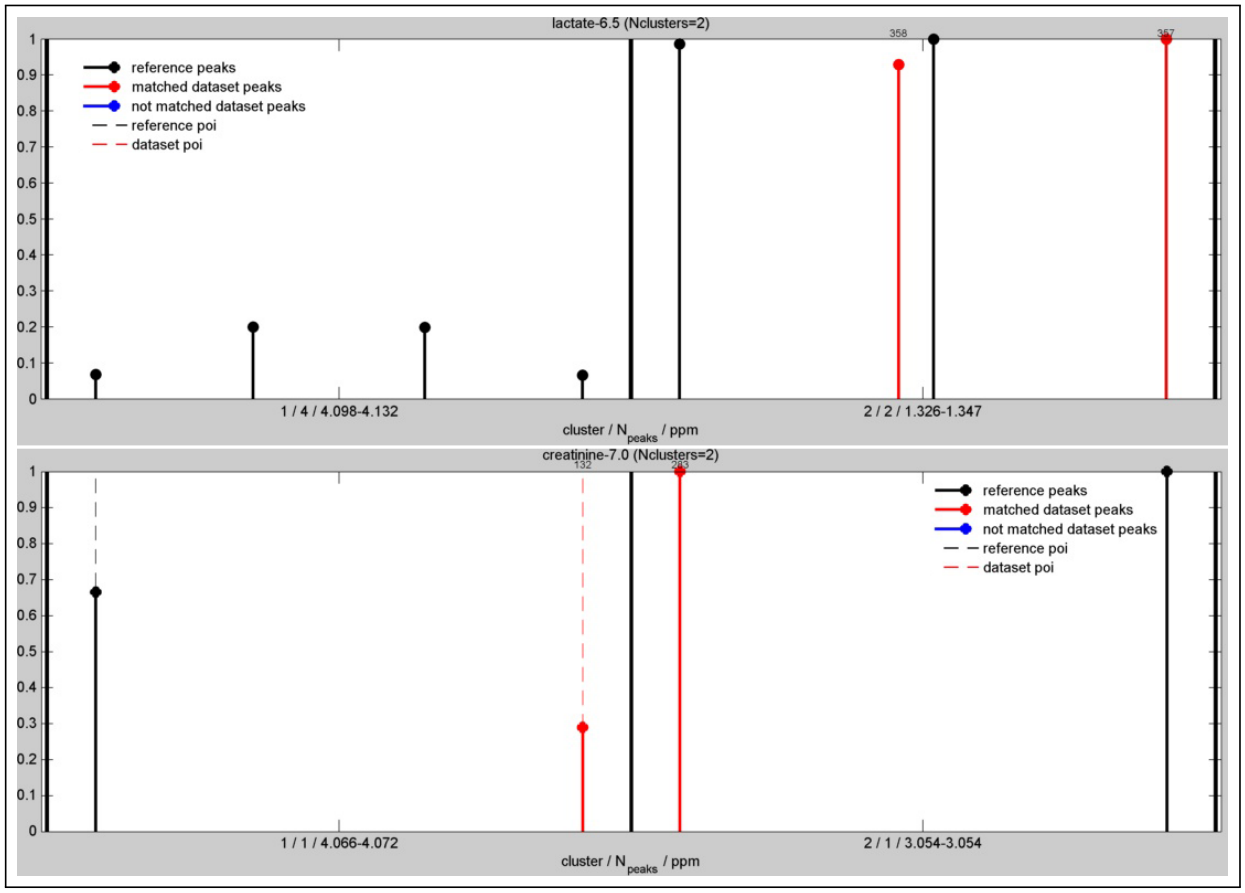


Figure B.13: *Metabolite identification on the urine dataset.* This figure shows two examples of successful identifications on the urine dataset (i.e. Lactate and Creatinine).

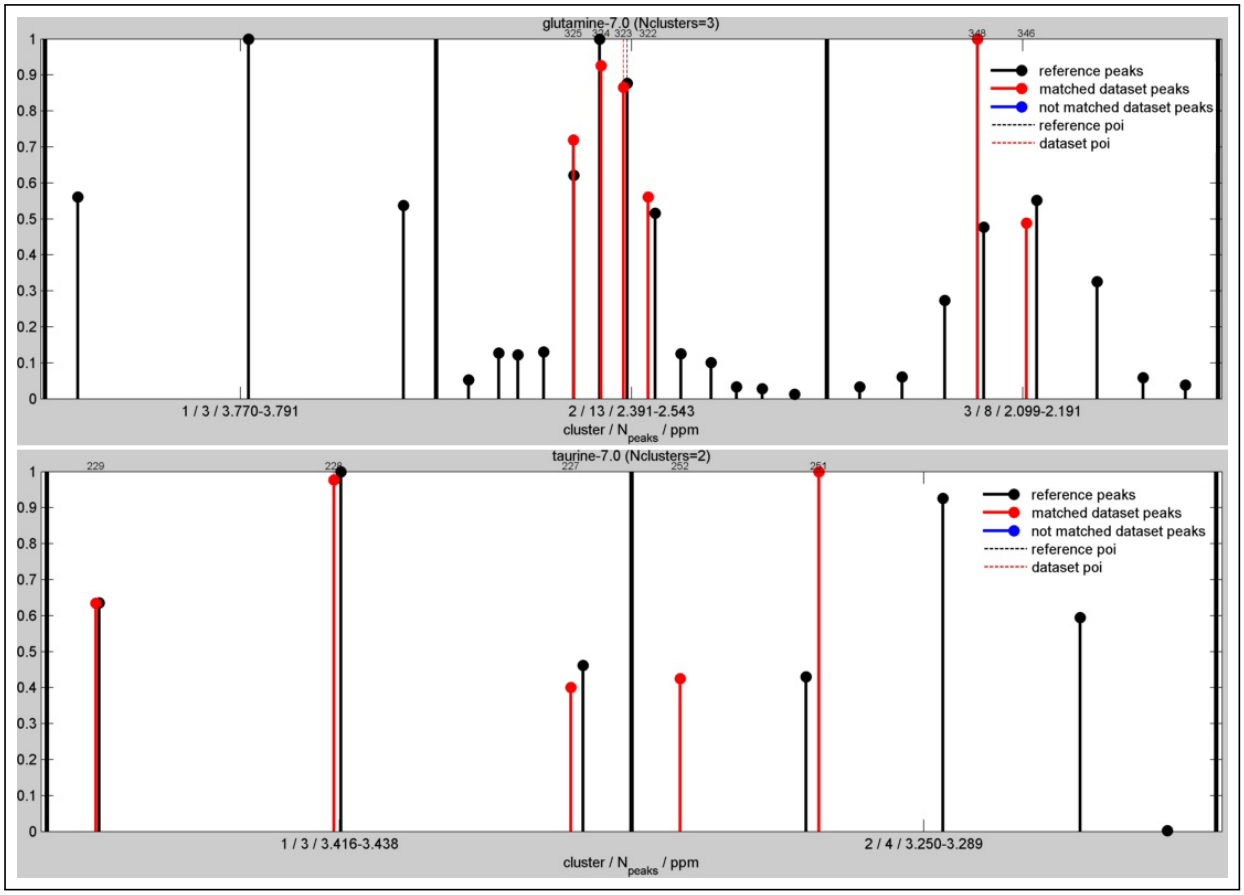


Figure B.14: *Metabolite identification on the liver extracts dataset.* This figure shows two examples of successful identifications (i.e.Glutamine and Taurine) on the liver extracts dataset.

B.2 Supplementary tables

METABOLITE	DATASET	METABOLITE	DATASET
β-hidroxyisovalerate	Urine	Pseudouridine	Urine
3-Aminoisobutyrate	Urine	Pyroglutamate	Urine
3-Methylhistidine	Urine	Serine	Urine
4-Hydroxyphenylacetic	Urine	Succinate	Urine
7-Methylxantine	Urine	TMAO	Urine
Acetone	Urine	Trigonelline	Urine
Acetylcarnitine	Urine	UMP	Liver
ADP	Liver	Uridine	Liver
Ascorbate	Liver	Valine	Liver
Betaine	Urine	β-hidroxybutyrate	Liver/Urine
Carnitine	Liver	Acetate	Liver/Urine
Cis-Aconitate	Urine	Alanine	Liver/Urine
Citrate	Urine	Choline	Liver/Urine
Dimethylamine	Urine	Creatine	Liver/Urine
Ethanol	Urine	Creatinine	Liver/Urine
Ethanolamine	Urine	Glucose	Liver/Urine
Formate	Urine	Glutamate	Liver/Urine
Fumarate	Liver	Glutamine	Liver/Urine
Glucose-D-1-Phosphate	Liver	Histidine	Liver/Urine
Glycerol	Liver	Isoleucine	Liver/Urine
Glycerophosphocholine	Urine	Lactate	Liver/Urine
Glycine	Urine	Leucine	Liver/Urine
Hippurate	Urine	Lysine	Liver/Urine
Indoxylsulfate	Urine	Phenylalanine	Liver/Urine
Methanol	Urine	Pyruvate	Liver/Urine
Methionine	Liver	Taurine	Liver/Urine
Methylguanidine	Urine	Threonine	Liver/Urine
NAD	Liver	Tyrosine	Liver/Urine
N-N-Dimethylglycine	Urine	Uracil	Liver/Urine

Table B.1: Metabolite databases used for identification. This table shows the metabolites included on the databases used for the liver extract and human urine identification analyses.

Peak	$\delta(ppm)$	Compound	C::P[1]	Score	Peak	$\delta(ppm)$	Compound	C::P[1]	Score
350	1.931	Acetate	1::1	0.68	219	3.577	Glycine	1::1	0.83
331	2.241	Acetone	1::1	0.88	19	7.846	Hippurate	1::1	1.00
269	3.201	Acetylcarnitine	4::1	0.86	22	7.832	Hippurate	1::2	1.00
354	1.496	Alanine	2::1	0.95	28	7.659	Hippurate	2::1	1.00
355	1.484	Alanine	2::2	0.95	29	7.647	Hippurate	2::2	1.00
268	3.209	Choline	3::1	0.96	33	7.634	Hippurate	2::3	1.00
281	3.107	Cis-Aconitate	2::1	0.74	36	7.571	Hippurate	3::1	0.98
303	2.692	Citrate	1::1	1.00	37	7.558	Hippurate	3::2	0.98
304	2.667	Citrate	1::2	1.00	38	7.546	Hippurate	3::3	0.98
306	2.552	Citrate	2::1	1.00	148	3.958	Hippurate	4::1	0.98
308	2.527	Citrate	2::2	1.00	357	1.347	Lactate	2::1	0.94
157	3.939	Creatine	1::1	0.96	358	1.335	Lactate	2::2	0.94
287	3.032	Creatine	2::1	0.95	242	3.367	Methanol	1::1	0.93
283	3.054	Creatinine	2::1	1.00	296	2.832	Methylguanidine	1::1	0.78
132	4.072	Creatinine	1::1	0.97	46	7.430	Phenylalanine	1::2	0.66
299	2.720	Dimethylamine	1::1	0.97	47	7.419	Phenylalanine	1::3	0.66
202	3.684	Ethanol	1::1	0.99	325	2.348	Pyruvate	1::1	0.74
373	1.203	Ethanol	2::1	0.93	240	3.418	Taurine	1::2	0.92
377	1.190	Ethanol	2::2	0.93	257	3.288	TMAO	1::1	0.85
378	1.178	Ethanol	2::3	0.93	117	4.448	Trigonelline	7::2	0.97
206	3.671	Ethanol	1::2	0.92	4	9.130	Trigonelline	2::1	0.96
207	3.660	Ethanol	1::3	0.92	6	8.843	Trigonelline	4::4	0.95
208	3.648	Ethanol	1::4	0.92	7	8.853	Trigonelline	4::3	0.95
9	8.468	Formate	1::1	0.93					

[1] Metabolite reference cluster and peak that have been associated with the dataset peak.

Table B.2: Peak identification on the human urine dataset. This table shows the dataset peaks correctly associated with a metabolite. C::P refers to the

Peak	$\delta(ppm)$	Compound	C::P[1]	Score	Peak	$\delta(ppm)$	Compound	C::P[1]	Score	Peak	$\delta(ppm)$	Compound	C::P[1]	Score
361	1.920	Acetate	1::1	0.99	325	2.440	Glutamine	2::9	0.84	367	1.728	Lysine	4::5	0.83
14	8.233	ADP	2::1	0.80	324	2.453	Glutamine	2::8	0.83	368	1.716	Lysine	4::6	0.83
371	1.489	Alanine	2::1	0.95	184	3.647	Glycerol	2::3	0.90	22	7.432	Phenylalanine	1::2	0.78
373	1.477	Alanine	2::2	0.95	389	1.010	Isoleucine	5::2	0.93	23	7.420	Phenylalanine	1::3	0.78
381	1.208	β-Hidroxybutyrate	4::1	0.88	386	1.022	Isoleucine	5::1	0.91	25	7.340	Phenylalanine	1::7	0.78
383	1.197	β-Hidroxybutyrate	4::2	0.88	95	4.118	Lactate	1::2	1.00	26	7.327	Phenylalanine	1::8	0.78
264	3.206	Choline	3::1	0.97	96	4.128	Lactate	1::1	1.00	227	3.437	Taurine	1::1	1.00
212	3.515	Choline	2::1	0.77	98	4.105	Lactate	1::3	1.00	228	3.427	Taurine	1::2	1.00
275	3.041	Creatine	2::1	1.00	99	4.094	Lactate	1::4	1.00	229	3.416	Taurine	1::3	1.00
274	3.057	Creatinine	2::1	0.85	376	1.336	Lactate	2::1	1.00	251	3.261	Taurine	2::3	0.93
38	6.521	Fumarate	1::1	0.97	377	1.325	Lactate	2::2	1.00	252	3.250	Taurine	2::4	0.93
73	4.642	Glucose	2::2	1.00	392	0.978	Leucine	3::1	0.95	30	7.205	Tyrosine	1::2	0.92
61	5.238	Glucose	1::1	0.95	396	0.966	Leucine	3::2	0.95	31	7.191	Tyrosine	1::3	0.92
62	5.231	Glucose	1::2	0.95	397	0.956	Leucine	3::3	0.95	34	6.909	Tyrosine	2::1	0.92
72	4.654	Glucose	2::1	0.95	366	1.742	Leucine	2::6	0.85	35	6.895	Tyrosine	2::2	0.92
162	3.739	Glucose	3::11	0.91	370	1.705	Leucine	2::10	0.85	18	7.871	Uridine	1::2	0.85
165	3.709	Glucose	3::13	0.91	368	1.716	Leucine	2::9	0.84	19	7.884	Uridine	1::1	0.85
127	3.888	Glucose	3::2	0.90	363	1.754	Leucine	2::5	0.82	385	1.041	Valine	3::2	0.90
163	3.730	Glucose	3::12	0.90	367	1.728	Leucine	2::8	0.82	394	0.989	Valine	3::4	0.90
334	2.360	Glutamate	2::5	0.73	363	1.754	Lysine	4::3	0.83	384	1.053	Valine	3::1	0.88
323	2.463	Glutamine	2::7	0.86	366	1.742	Lysine	4::4	0.83	393	1.000	Valine	3::3	0.88

[1] Metabolite reference cluster and peak that have been associated with the dataset peak.

Table B.3: Peak identifications on the liver extracts dataset. This table shows the dataset peaks correctly associated with a metabolite. C::P refers to the metabolite reference cluster and peak that have been associated with the dataset peak.

B.3 FOCUS methodology

B.3.1 Spectral segmentation

FOCUS algorithm performs spectral segmentation in order to independently align consecutive segments within the spectra in a unsupervised way (see Figure B.1). These segments are defined with a certain degree of overlap (i.e. 50%) so peaks near one segment edge will be centered in the adjacent segment. A counterpart of this approach is the possibility of having the same peak repeated across consecutive segments. Nevertheless, after the alignment and peak picking steps, FOCUS introduces a peak reduction method that takes this into account to remove redundant peaks.

B.3.2 RUNAS alignment algorithm

Intensity-weighted signal transform

Previous to spectral alignment, FOCUS applies a mathematical transformation in order to improve posterior alignment results (see Figure B.2). This transformation is applied to each input spectrum segment and proceeds as follows:

1. **Spectral segment weighting:** Since FOCUS algorithm is aimed to be applied iteratively within overlapping segments of the entire spectrum, each segment is previously weighted in order to reduce the weight of the segment points near the edges. Therefore, each spectral segment is multiplied by a Tuckey window defined with a ratio of taper to constant sections of 0.4 when segment overlap is set to 50%.

2. **Moving average filtering:** The resulting segment signal from the previous step is filtered with a moving average filter in order to avoid spurious sign changes in the signal derivative.

3. **Signal derivative sign:** The next step consists of computing the signal derivative sign as follows,

$$\bar{s}_i^t = \frac{\partial s_i^t}{\partial x} \tag{B.1}$$

   where s_i^t is the spectrum segment t of sample i once steps 1 and 2 have been applied.

4. **Percentile values and limits:** Given a number of percentile quantification values (usually $N_Q = 10$) the percentile values q_i and the signal limits L_i for each percentile value are defined as follows,

$$q_i = 100\frac{i-1}{N_Q - 1}, \forall i \in [1, N_Q] \tag{B.2}$$

$$L_i = perc(\bar{s}_i^t, i), \forall i \in [1, N_Q] \tag{B.3}$$

where $perc(x, p)$ refers to the percentile p of signal x.

5. **Percentile weighting of the signal derivative sign:** The final step for transforming the input spectrum segment applies the following formula at each segment point,

$$S_i^t(x) = \bar{s}_i^t \sum_{i=1}^{N_Q} q_i I(L_{i-1} < \bar{s}_i^t \leq L_i) \tag{B.4}$$

where $I(g)$ is the indicator function which value is 0 until the condition inside is satisfied, in which case is set to one.

The resulting signal has multiple advantages for spectral alignment:

- increasing and decreasing spectral regions have opposite signs and, consequently, the correlation function will be more sensitive to unalignments within peak positions
- the correlation function will also tend to match segments with the same slope sign
- the signal dynamic range (i.e. signal intensity differences between different peaks) is reduced avoiding high contrasts on the correlation contribution of large versus small intensity peaks
- value ranges across the different spectra are equalized.

Recursive unreferenced alignment

Given the following set of transformed segment spectra,

$$S_i^t(x) \begin{cases} i \in [1, N_S] \\ t \equiv \text{spectral segment being aligned} \end{cases} \tag{B.5}$$

FOCUS alignment algorithm is applied in two steps (see Figure B.3):

1. **MCM and ODM calculation:** MCM (Maximal Correlation Matrix) stores the maximal correlation that can be achieved between each pair of segment spectra considering all the possible shifts. ODM (Optimal Distance Matrix) stores the shift to be applied to each segment in order to achieve the maximal correlation with another sample segment. These matrices are computed as follows,

$$ODM_{ij}^0 = \max_{\delta \in [1, L_S]} (C(S_i^t(x - \delta), S_j^t(x))) \tag{B.6}$$

$$MCM_{ij} = C(S_i^t(x - ODM_{ij}), S_j^t(x))) \tag{B.7}$$

where i and j are the pair of sample segments considered, L_S is the segment length, δ the segment shift and $C(g)$ the correlation function. FFT is used in order to speed up this computation.

2. **Recursive alignment:** Recursive alignment is based on the previous matrices and on a correlation threshold that excludes from the shift calculation of each segment all the other segments that do not achieve a maximal correlation above this threshold. Four steps are recursively iterated until convergence:

   (a) Shift calculation for each segment:

$$\delta_i^k = \frac{\sum_{j=1}^{N_S} I(MCM_{ij} \geq C_T) MCM_{ij} ODM_{ij}^k}{2 \sum_{j=1}^{N_S} I(MCM_{ij} \geq C_T) MCM_{ij}}, \forall i \in [1, N_S] \tag{B.8}$$

   where k refers to the algorithm iteration and N_S to the number of segments to be aligned.

   (b) Segment shift update:

$$S_i^t(x) = S_i^t(x - \delta_i^k), \forall i \in [1, N_S] \tag{B.9}$$

   (c) ODM update:

$$ODM_{ij}^{k+1} = ODM_{ij}^k - \delta_i^k + \delta_j^k \tag{B.10}$$

   (d) Convergence check: Recursive alignment stops either if the applied shifts to all the spectral segments are 0 or if the number of maximal iterations has been achieved:

$$\text{stop conditions} \begin{cases} \sum_{i=1}^{N_S} I(\delta_i^k = 0) = N_S \\ k > K_{max} \end{cases} \tag{B.11}$$

B.3.3 Peak detection

Peak picking and area integration is performed by FOCUS using as input data the aligned segment spectra defined as follows:

$$s_i(n) \begin{cases} i \in [1, N_S] \\ n \in [1, N_P] \end{cases} \tag{B.12}$$

where N_S is the number of sample spectra and N_P the number of spectral points within the analyzed segment.

First step consists of a per-sample peak identification and the calculation of a consensus peak signal $p(n)$ that stores at each segment point the global peak frequency across the samples (see Figure B.4). This signal is computed as follows:

1. The consensus peak signal is initialized to zero: $p(n) = 0, \forall n$
2. Peak identification at the sample spectrum level is performed as follows:
   (a) The spectrum i is first filtered using a second derivative gaussian signal $\Psi_{LF}(n)$:

$$s_i^f(n) = s_i(n)\Psi_{LF}(n) \tag{B.13}$$

   where LF refers to the filter length and its default value is 0.01 ppm.

   (b) Once filtered, FOCUS obtains for each sample the zero-crossing signal $d_i^f(n)$ derived from $s_i^f(n)$:

$$d_i^f(n) = I(s_i^f(n) > 0) - I(s_i^f(n-1) > 0) \tag{B.14}$$

   where $I(g)$ refers to the indicator function. The resulting signal will worth 0 unless a negative to positive (+1) or positive to negative (-1) zero-crossings are found.

   (c) For each sample, the "peak regions" are identified and delimited over $d_i^f(n)$. The peak region j delimited by points a_{ij} and b_{ij} must fulfill the following six criteria:

$$(1) \Rightarrow a_{ij} < b_{ij} \tag{B.15}$$
$$(2) \Rightarrow d_i^f(a_{ij}) = 1 \tag{B.16}$$
$$(3) \Rightarrow d_i^f(b_{ij}) = -1 \tag{B.17}$$
$$(4) \Rightarrow d_i^f(n) = 0, \forall n \in (a_{ij}, b_{ij}) \tag{B.18}$$
$$(5) \Rightarrow \max_{n\in(a_{ij},b_{ij})}(s_i^n) - \min_{n\in(a_{ij},b_{ij})}(s_i^n) \geq T_I^1 \tag{B.19}$$
$$(6) \Rightarrow \frac{\max_{n\in(a_{ij},b_{ij})}(s_i^n)}{\min_{n\in(a_{ij},b_{ij})}(s_i^n)} \geq T_I^2 \tag{B.20}$$

   where i refers to the sample spectrum, j identifies the peak region, $a_{ij} - b_{ij}$ delimit the peak region, and T_I^1 and T_I^2 respectively refer to the minimum intensity increment and minimum percentage increment in a region to be considered as a peak.

3. Once FOCUS has computed the "peak regions" for all the sample spectra, the consensus peak signal is easily updated as follows:

$$p(a_{ij} \leq nb_{ij}) = p(a_{ij} \leq nb_{ij}) + 1 \begin{cases} \forall i \in [1, N_S] \\ \forall j \in [1, N_{PR}^i] \end{cases} \tag{B.21}$$

where N_{PR}^i refers to the number of "peak regions" identified over the spectrum i.

4. Finally, the consensus peak signal is scaled and filtered with a moving rectangular window $\Pi_{LF/4}$ of length $LF/4$ where LF refers to the filter length used in equation B.13:

$$p(n) = \frac{p(n)}{N_S} \Pi_{LF/4}(n) \tag{B.22}$$

Once the consensus peak signal has been computed the overall segment peak regions can be easily computed. The value of this consensus peak signal can be regarded as the peak frequency on the analyzed sample spectra (i.e. $p(n) = 1$ means that all the samples have a peak region spanning the spectral point n). Thus, the global peak regions that will define a peak are identified by repeating steps 2a to 2c only over the consensus peak signal. Criteria to consider a peak on the consensus peak signal are defined as follows:

$$(1) \Rightarrow a_{ij} < b_{ij} \tag{B.23}$$

$$(2) \Rightarrow d_i^f(a_{ij}) = 1 \tag{B.24}$$

$$(3) \Rightarrow d_i^f(b_{ij}) = -1 \tag{B.25}$$

$$(4) \Rightarrow d_i^f(n) = 0, \forall n \in (a_{ij}, b_{ij}) \tag{B.26}$$

$$(5) \Rightarrow \max_{n \in (a_{ij}, b_{ij})} (p(n)) \geq f_T \tag{B.27}$$

Finally, the global peak regions can be expanded or rejected following these two criteria:

1. If the peak region is narrower than a predefined minimal peak width (which its default value is set to $1e^{-3}$ ppm) it is rejected.

2. The peak region is expanded to a maximum range value of 0.5 times its width unless another peak region is found or the consensus peak signal reaches a zero.

B.3.4 Peak reduction

Before proceeding with metabolite identification, the list of peaks must be analyzed in order to reduce their redundancy. This peak reduction method is applied to reject redundant

peaks that have been obtained in consecutive spectral segments. This is specifically intended for unsupervised analysis where the spectra are divided in consecutive segments with a 50% overlap. Thus, a peak can be obtained twice on two consecutive segments and this redundancy must be avoided.

Considering two consecutive overlapping segments i and i+1 the peak reduction method takes into account all the peaks detected within each segment:

$$p_i^m \Rightarrow m \in [1, N_i] \tag{B.28}$$

$$p_{i+1}^m \Rightarrow m \in [1, N_{i+1}] \tag{B.29}$$

where p_i^m and p_{i+1}^n respectively refer to the peaks detected in segments i and $i+1$. N_i and N_{i+1} are the total number of peaks detected on each segment.

For each pair of peaks (p_i^m, p_{i+1}^n FOCUS computes the average per-sample cross-peak overlap:

$$O_{m,n}^s = \frac{\max(0, \min(x_R^s(p_i^m), x_R^s(p_{i+1}^n)) - \max(x_L^s(p_i^m), x_L^s(p_i^m), x_L^s(p_{i+1}^n)))}{\max(x_R^s(p_i^m) - x_L^s(p_i^m), x_R^s(p_{i+1}^n) - x_L^s(p_{i+1}^n))} \tag{B.30}$$

$$O_{m,n} = \frac{1}{N_S} \sum_{s=1}^{N_S} O_{m,n}^s \tag{B.31}$$

where $O_{m,n}^s$ is the spectral overlap between peaks (p_i^m) and (p_{i+1}^n) for sample s and $O_{m,n}$ the average per-sample cross-peak overlap. $x_R^s(p)$ and $x_L^s(p)$ respectively refer to the right and left limits of peak p defined for sample s.

Peak pairs having $O_{m,n}^s \geq 0.5$ are selected and the method keeps only the peak that has a greater peak shape correlation.

B.3.5 Metabolite identification

Metabolite identification is based on matching reference spectra corresponding to each metabolite to the set of peaks identified on the analyzed data. A metabolite reference spectrum is defined by its spectral peaks, which are defined by its positions (ppm) and relative intensities (i.e. heights):

$$M_i \equiv \{p_i^r = \left[\widetilde{x}_i^r, \widetilde{I}_i^r\right]\}, r \in [1 \ldots R_i] \tag{B.32}$$

where metabolite $i(M_i)$ is defined by R_i peaks. $\widetilde{x}_i^r$ refers to the reference peak position and $\widetilde{I}_i^r$ to its relative intensity.

Before starting the identification of each dataset peak, FOCUS clusterizes each reference metabolite spectrum obtaining a set of peak reference clusters that group close spectrum

peaks:

$$|\widetilde{x}_i^r - \widetilde{x}_i^s| < d_{clust} \quad \Rightarrow \quad C_i^r = C_i^s \tag{B.33}$$

$$|\widetilde{x}_i^r - \widetilde{x}_i^s| \geq d_{clust} \quad \Rightarrow \quad C_i^r \neq C_i^s \tag{B.34}$$

$$\tag{B.35}$$

where C_i^r refers to the cluster identifier of the reference peak r within metabolite i.

Once all the metabolite spectra have been clusterized, the identification process for each dataset peak begins. Given a dataset peak:

$$D_p \begin{cases} x_p \equiv \text{Dataset peak position (ppm)} \\ I_p^s \equiv \text{Dataset peak intensity for sample s} \\ I_p^c = \text{median}(I_p^1 \dots I_p^s \dots I_p^{N_p}) \equiv \text{Dataset peak median intensity} \end{cases} \tag{B.36}$$

it will be matched against all the reference metabolites that have, at least, one peak close to the dataset peak:

$$|x_p - \widetilde{x}_i^r| \leq t_{cluster} \Rightarrow M_i \in CAND_p \tag{B.37}$$

where $CAND_p$ refers to the set of metabolite candidates for the dataset peak p.

Given one dataset peak D_p (equation B.36) and one matched peak r_i from a candidate metabolite M_i (equation B.32) their identification score is computed as follows:

- If the candidate metabolite has only one spectral peak, the identification score will be based on averaging the three following components:

  1. The first component score is based on the distance between the reference peak and the dataset peak. This score is scaled with respect to twice the cluster tolerance:

  $$S_1 = 1 - \frac{\min\left(|x_p - \widetilde{x}_i^r|, 2t_{cluster}\right)}{2t_{cluster}} \tag{B.38}$$

  2. The second component takes into account the peak quality score computed as the per-sample average correlation of peak shapes:

  $$S_2 = Q_P \tag{B.39}$$

  3. The last component consists of the penalization score introduced to penalize the presence of other dataset peaks highly correlated with the matched dataset peak:

  $$S_3 = \max(0, 1 - wN) \tag{B.40}$$

where w refers to a weighting factor and N to the number of highly correlated dataset peaks.

- If the dataset peak has multiple spectral peaks, the identification score is based on the three following components:

  1. Intra-cluster component: If the peak reference cluster C_i has more than one peak, a set of search windows are defined around each reference peak that belongs to the same cluster than the matched reference peak. These search windows are placed in a way to maintain the position pattern of the reference intra-cluster peaks, taking also into account the distance between the matched reference and dataset peaks. Left and right window limits are computed as follows:

$$\forall m \in C_i^r \Rightarrow \begin{cases} L_c^m = x_p - (\widetilde{x}_i^r - \widetilde{x}_i^m) - \frac{t_{peak}}{2} \\ R_c^m = x_p - (\widetilde{x}_i^r - \widetilde{x}_i^m) + \frac{t_{peak}}{2} \end{cases} \tag{B.41}$$

  Once these windows have been defined, the algorithm searches for all the dataset peaks spanning them and related with the peak that is being identified by a per-sample intensity correlation coefficient exceeding a correlation threshold. If a search window does not identify any related dataset peak a zero-intensity peak is assigned (i.e. it has not been found). Finally, the correlation between the intensity levels of the related dataset peaks and their respective reference peaks is computed in order to obtain the intra-cluster matching score:

$$S_1 = \text{corrcoef}_{C_i^r = m, r \leftrightarrow p}\{\widetilde{I}_i^r, \widetilde{I}_p^c\} \tag{B.42}$$

  where m refers to the cluster identifier of the matched reference peak and $r \leftrightarrow p$ means that dataset peak D_p (equation B.36) and reference peak p_i^r (equation B.32) have been matched.

  2. Inter-cluster component is based on the number of reference metabolite clusters having correlated dataset peaks. Inter-cluster search windows are defined to maintain the position distances between the reference peak matched and the reference peaks of the other clusters. A tolerance window using $t_{cluster}$ is also defined across the first and the last peak of each cluster. The final inter-cluster score is computed as follows:

$$S_2 = \frac{\sum\limits_{m \in C_i} F_m \max_{C_i^r = m}(\widetilde{I}_i^r)}{\sum\limits_{m \in C_i} \max_{C_i^r = m}(\widetilde{I}_i^r)} \tag{B.43}$$

where C_i refers to the set of clusters identified for metabolite i and F_m is set to 0 unless a correlated dataset peak has been found within the search window for cluster m.

3. The last component consists of the penalization score introduced to penalize the presence of other dataset peaks highly correlated with the matched dataset peak. Correlated dataset peaks outside the metabolite cluster regions are taken into account for this calculation (see equation B.40).

Once all the three components have been computed, the final identification score for a candidate metabolite against a dataset peak is calculated as the average of the three components.

B.4 Alignment evaluation

B.4.1 Synthetic spectral datasets

Synthetic spectral dataset generation

In order to test the different alignment methods and both evaluate and compare their performance two synthetic datasets have been generated. These datasets are characterized by the presence of two and three peaks per sample respectively. The sample generated spectra have been derived from the following formulas using the Lorentzian probability density function (PDF) as the basis function for peak definition:

$$Doublet \Rightarrow y_s(x) = \frac{M_{s,1}(a+A_{s,1})}{\pi\lambda(1+(\frac{x-U_s+x_0}{\lambda})^2)} + \frac{M_{s,2}r(a+A_{s,2})}{\pi\lambda(1+(\frac{x-U_s-x_0}{\lambda})^2)} \quad \text{(B.44)}$$

$$Triplet \Rightarrow \frac{M_{s,1}(a+A_{s,1})}{\pi\lambda(1+(\frac{x-U_s+x_0}{\lambda})^2)} + \frac{M_{s,2}r(a+A_{s,2})}{\pi\lambda(1+(\frac{x-U_s}{\lambda})^2)} + \frac{M_{s,3}(a+A_{s,3})}{\pi\lambda(1+(\frac{x-U_s-x_0}{\lambda})^2)} \quad \text{(B.45)}$$

Spectra depend on deterministic (a, r, x_0, λ) and random variables (M, A, U) which are defined as follows:

- a: This parameter defines the mean peak intensity that is used for all the modeled peaks
- A: Normally distributed random variable with zero mean. The standard deviation of this variable defines the peak intensity variability across de samples.
- r: This parameter defines the intensity ratio between the generated peaks of each spectrum.

- M: Random binary variable ($[0,1]$) that accounts for the probability of a sample of having missing peaks. Frequency of zero values (missingness) is the parameter that controls this variable.
- λ: Scaling parameter of the Lorentzian PDF that accounts for the peak width.
- x_0: This parameter defines the distance between the different peaks in the generated spectra. This distance is $2x_0$ in the "doublet" dataset and x_0 in the "triplet" dataset.
- U: This random variable defines the simulated unalignment between samples. It follows a zero means normal distribution which standard deviation will be proportional to the sample unalignment degree.

In order to test alignment methods, each dataset has been simulated under a high number of scenarios defined by the parameters above. Since the generated spectral dataset depends on random variables, each scenario has been iterated n=10 times and the performance measurements averaged across them. In our simulations spectra have been defined with $N_X = 256$ data points, $a = 4$ and $\sigma(A_{s,i}) = 1.2$. The different scenarios have been obtained by permuting all the following parameter values:

- Distance between peaks: $x_0 \in [5, 10, 20, 40]$
- Scale parameter: $\in [2, 5]10$
- Peak intensity ratio: $\in [1, 2, 4]$
- Peak missingness: $f(M_{s,i} = 0) \in [0, 0.2, 0.5]$
- Standard deviation of unalignment shifts: $\sigma(U_s) \in [20, 30, 40]$
- Sample size: $N_S \in [25, 50, 100]$

Taking into account all the possible permutations, 972 scenarios were generated. Figure B.10 shows how the resulting spectral dataset changed when modifying one parameter while the others are set constants.

Alignment evaluation

For each simulated scenario all the true aligned spectra is known and saved before applying the artificial unalignment (U in the previous subsection). The ideal alignment method should therefore return back the spectra to their initial known positions. In order to provide correct performance measurements we have used two metrics, which are computed as follows:

1. **CORRELATION MATRIX DISTANCE**: Since the ultimate aim of the alignment methods is to recover the relative position that each spectrum had with respect to the others before applying unalignment, their performance is measured using the distance between the initial correlation matrix (previous to unalignment) and the final correlation matrix (after consecutively applying the simulated unalignment and the alignment method):

$$d_A = \sqrt{\frac{\sum_{i=1}^{N_S}\sum_{j=1}^{N_S}(C_{ij} - \hat{C}_{ij})^2}{N_S(N_S-1)}} \tag{B.46}$$

where A refers to the applied algorithm, C_{ij} to the correlation coefficient between spectra i and j before applying unalignment, $\hat{C}_{ij}$ to the correlation coefficient between spectra i and j before after applying alignment algorithm A and N_S to the number of sample spectra.

2. **SHIFT CORRELATION**: Since Focus and Icoshift are based on shifting each spectrum to maximize correlation their performance can be measured by comparing the correction shifts they apply with respect to the previously applied shifts to unalign the dataset. Therefore, this performance measurement uses correlation coefficient between applied shifts to unalign the dataset and the correction shifts applied by these two alignment methods.

B.4.2 Human urine spectral datasets

Alignment evaluation has also been tested over a complex dataset of 60 human urine spectra where unalignment effects of chemical shifts are clearly visible. 48 informative segments were selected across the ^{1}H-NMR spectra in order to base the performance measures on the alignment of relevant peaks (see Figure B.4).

The metric used for evaluation was the averaged spectra correlation matrix after and before applying the three alignment algorithms:

$$C_A = \frac{\sum_{i=1}^{N_S}\sum_{j=1}^{N_S} C_{ij}^A}{N_S(N_S-1)} \tag{B.47}$$

where N_S is the number of sample spectra and C_{ij}^A the correlation coefficient between sample spectra i and j. A can refer to spectra before alignment or after applying FOCUS, Icoshift or COW alignment.

Per-sample averaged spectra correlations were also computed in order to evaluate the alignment at the sample level:

$$C_A^i = \frac{\sum_{j=1, j\neq i}^{N_S} C_{ij}^A}{N_S(N_S-1)} \tag{B.48}$$

Finally, the number of sample spectra analyzed was varied to test the possible effect of sample size in alignment performance. Measures were taken for the whole dataset (N=60) but also for N=30 and N=10 samples. When evaluating the alignment over N=30 and N=10 the results were averaged across 20 sample permutations.

C | Supplementary Data of the Genome-Wide Association Study for Crohn's Disease Phenotypes

C.1 Supplementary figures

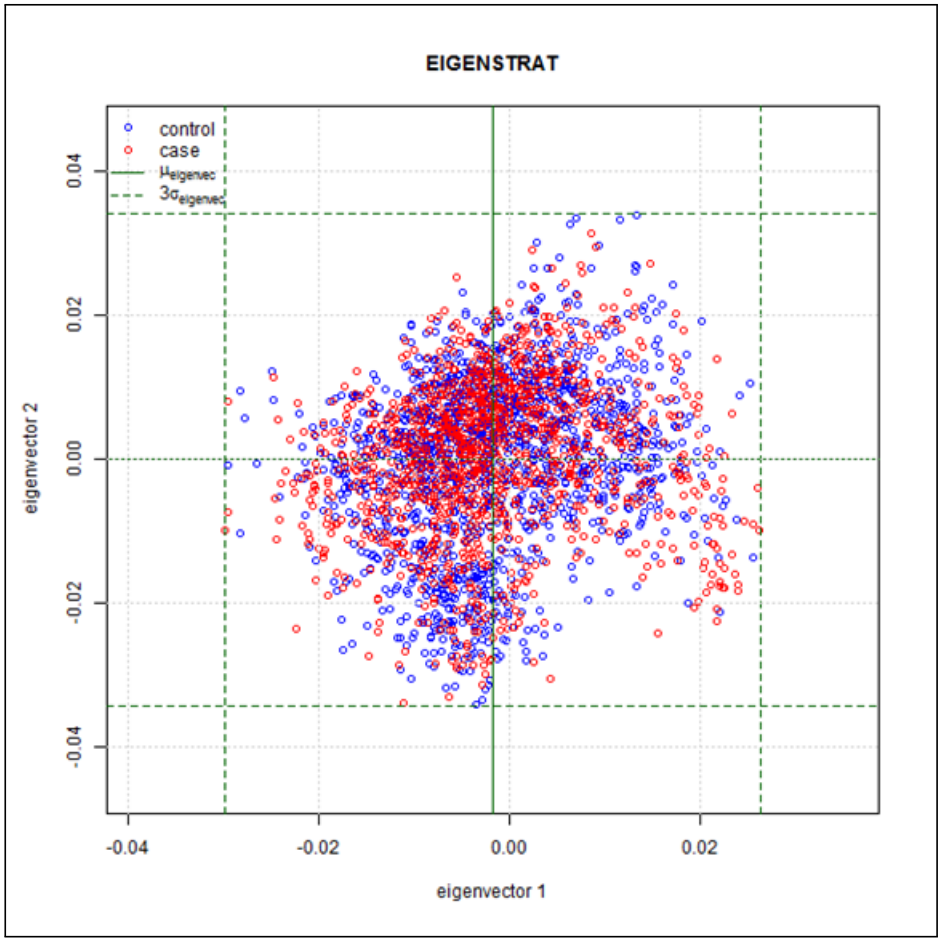


Figure C.1: *Distribution of principal components 1 and 2 between the control and CD samples included in the discovery GWAS analysis.* Both cohorts obtain very similar distributions thus discarding the need of population stratification adjustment in our analyses. The dashed green lines correspond to 3 times the standard deviation of the corresponding principal component scores. Samples outside these limits (n=76) were excluded (i.e. not shown).

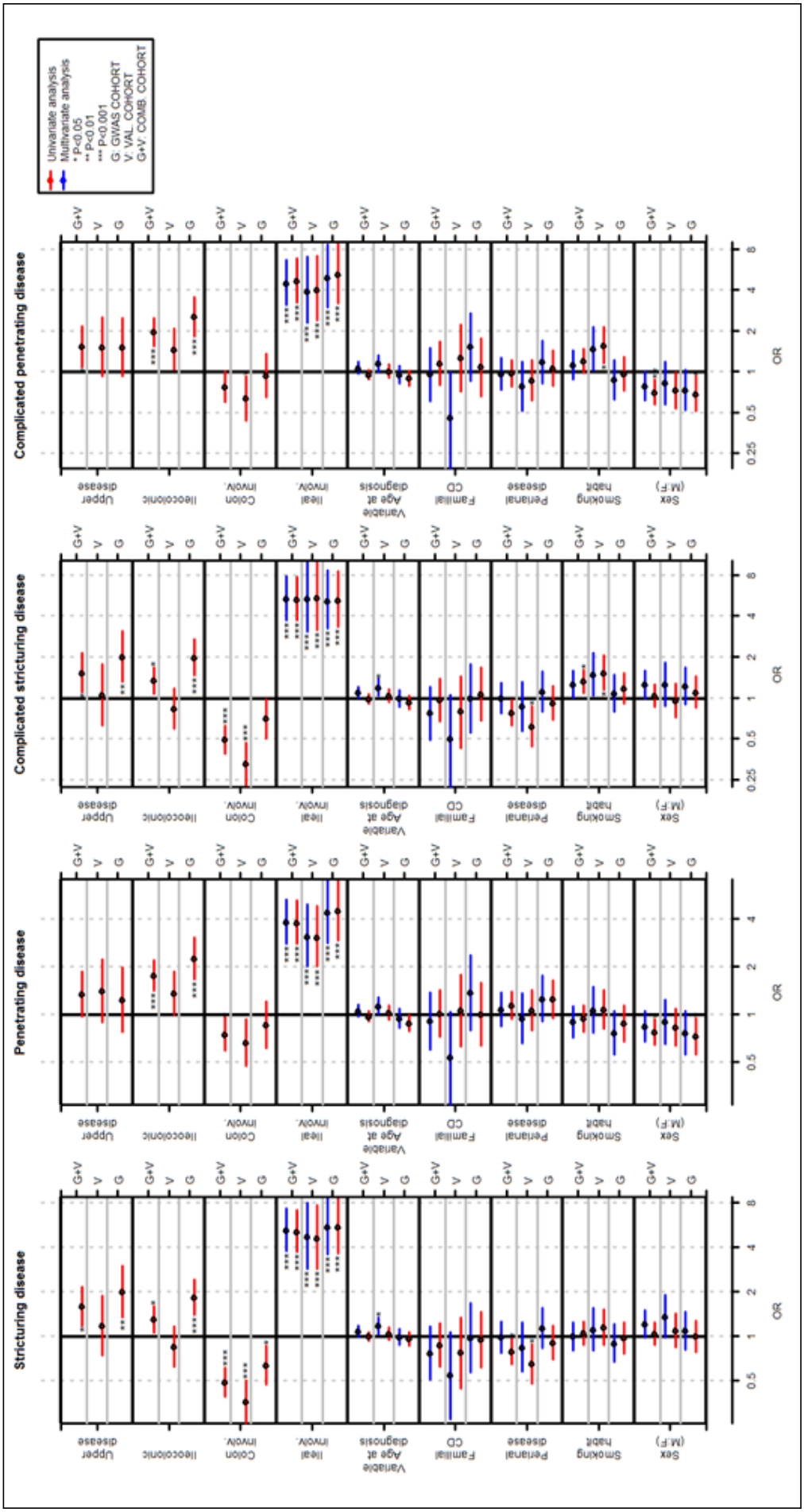


Figure C.2: *Logistic regression analysis of phenotypes related with disease behaviour.* This figure shows the dependence between the behaviour-related CD phenotypes and other clinical variables of interest. The ORs (95%CI) and significant associations of the logistic regression analysis are shown for each phenotype in the GWAS, validation and combined cohorts.

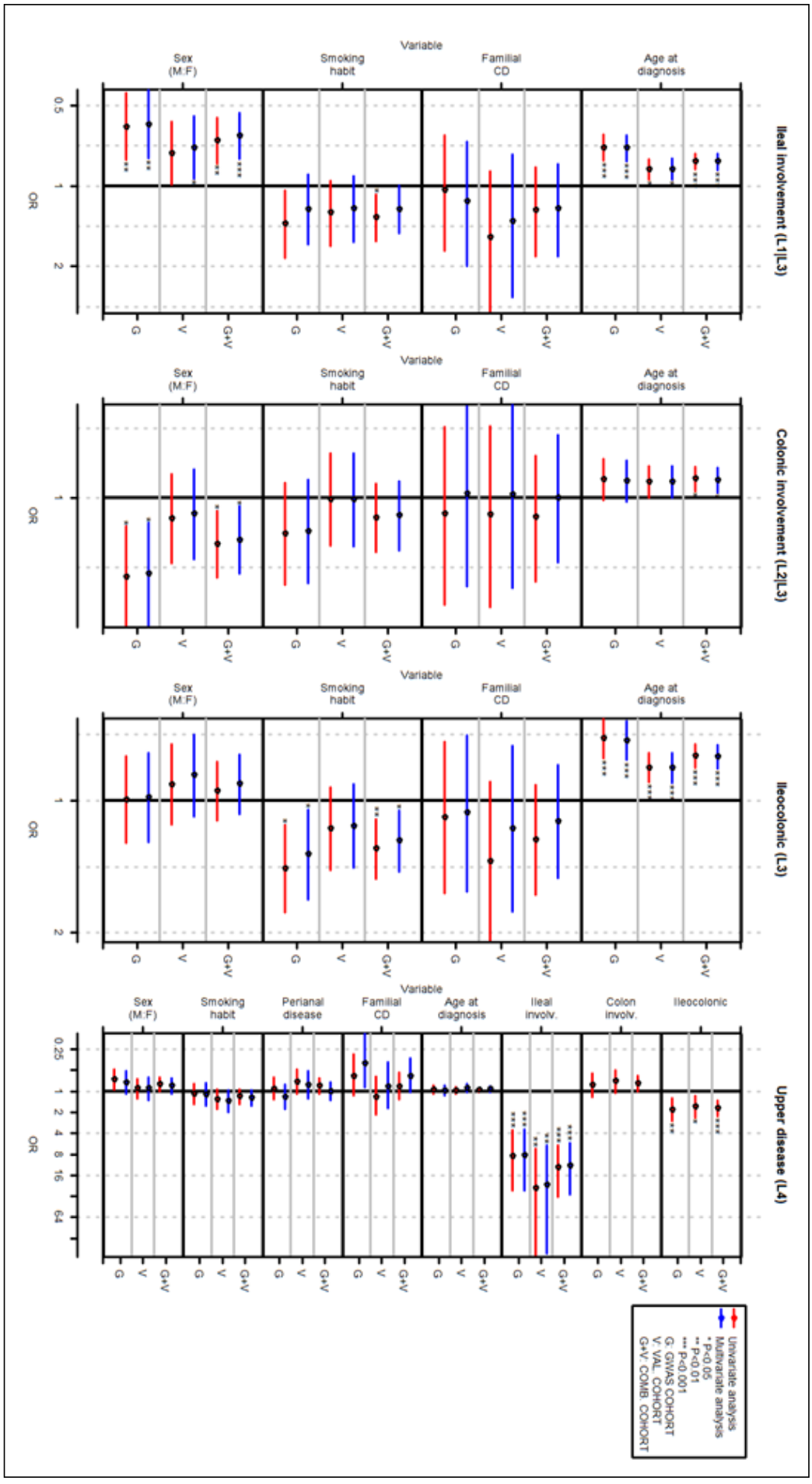


Figure C.3: *Logistic regression analysis of phenotypes related with disease location.* This figure shows the dependence between the location-related CD phenotypes and other clinical variables of interest. The ORs (95%CI) and significant associations of the logistic regression analysis are shown for each phenotype in the GWAS, validation and combined cohorts.

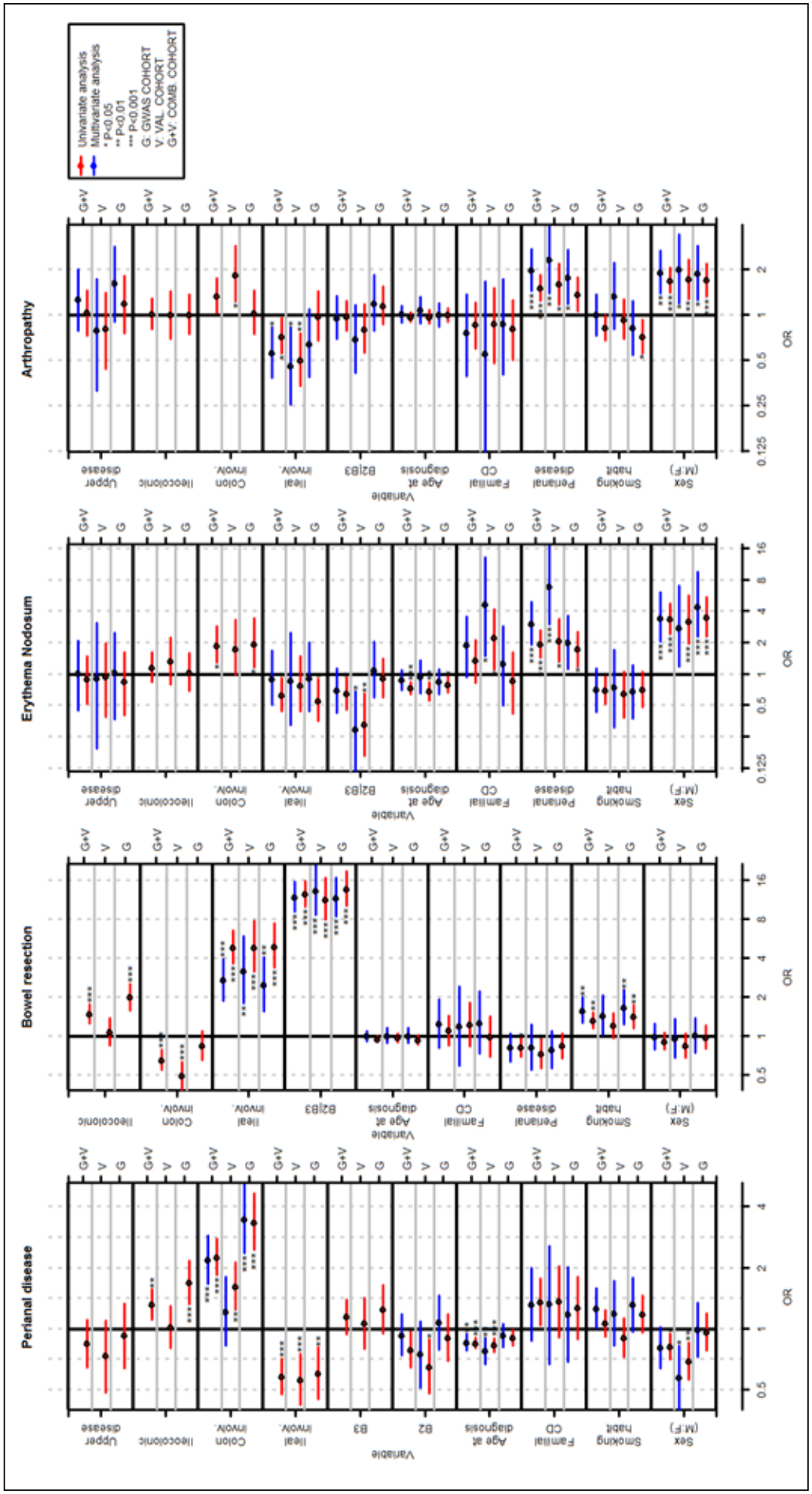


Figure C.4: *Logistic regression analysis of other CD phenotypes.* This figure shows the dependence between other CD phenotypes (i.e. perianal disease, bowel resection, erythema nodosum and arthropathy) and other clinical variables of interest. The ORs (95%CI) and significant associations of the logistic regression analysis are shown for each phenotype in the GWAS, validation and combined cohorts.

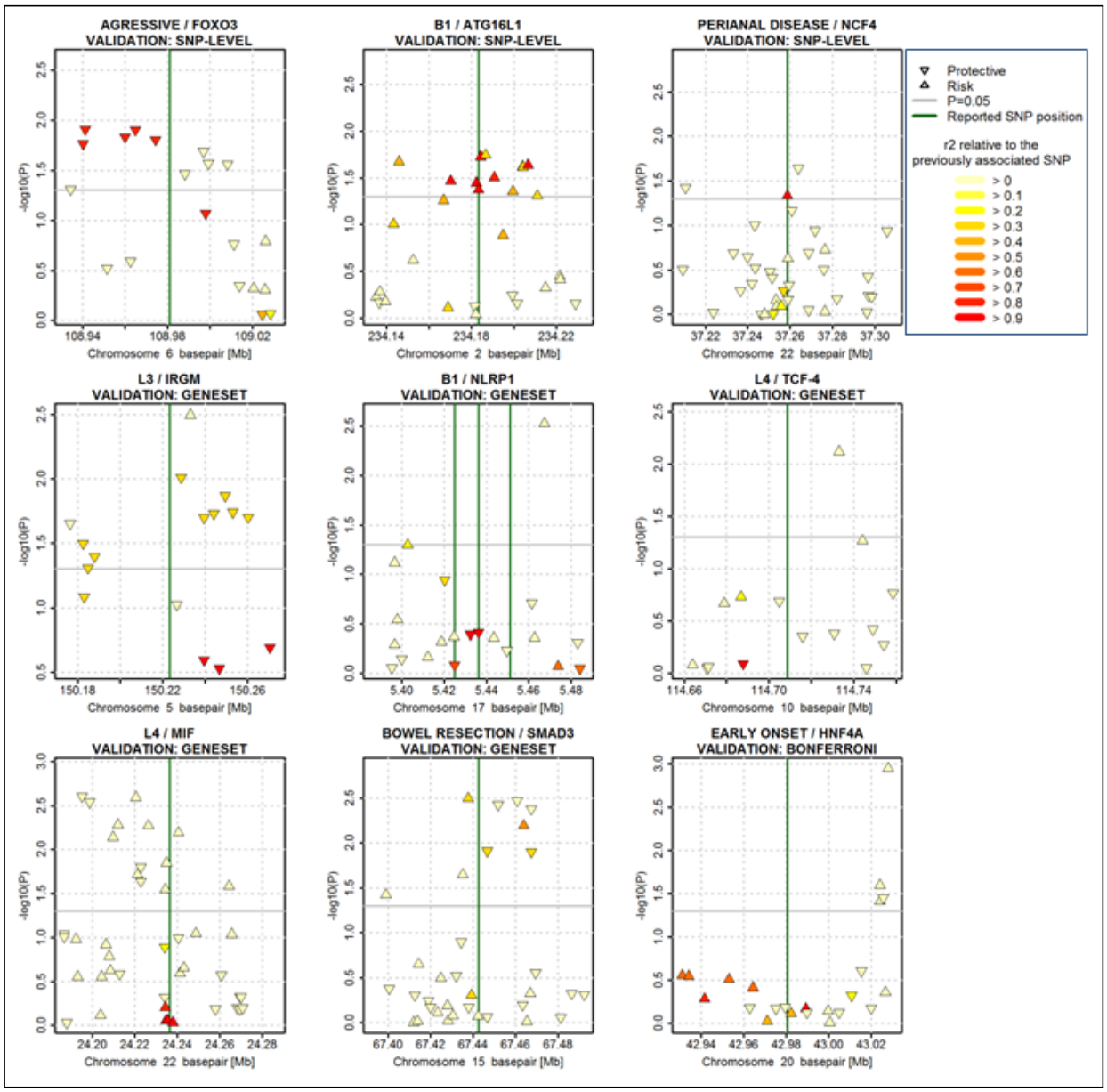


Figure C.5: *Association values at the previously reported phenotype susceptibility loci.* This figure shows the validated previously reported SNP-Phenotype associations. The marker color scale indicates the LD of the SNP regarding to the reported SNP (from 1 to 0).

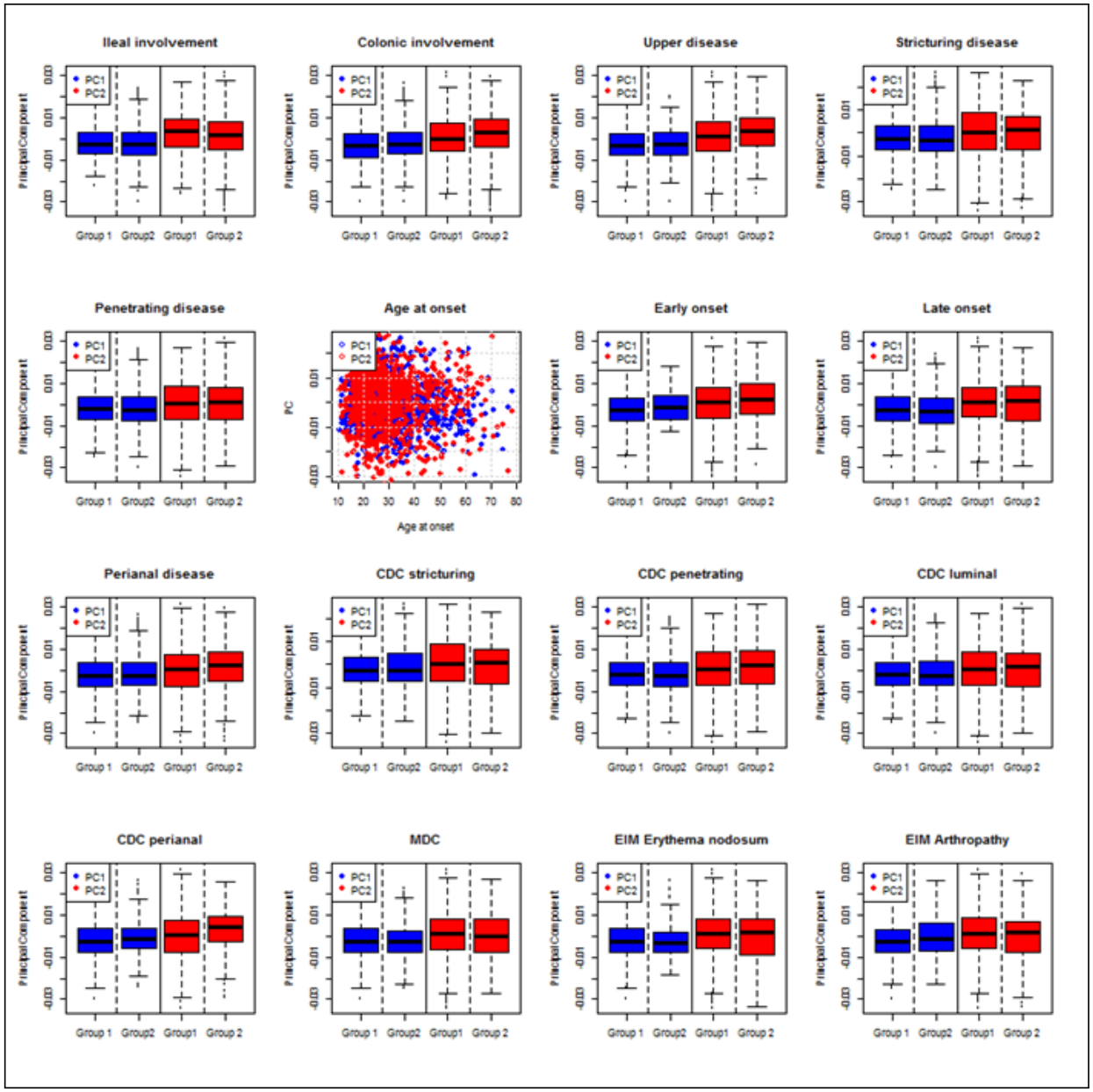


Figure C.6: *Distribution of principal components 1 and 2 within each phenotype to assess population stratification.* The boxplots show the distribution of principal components 1 and 2 within each discrete phenotype (groups 1 and 2 respectively refer to negative and positive groups as defined in Table 5.2). For continuous variables (e.g. age at onset), a scatter plot shows any relationship between principal components and the variable values.

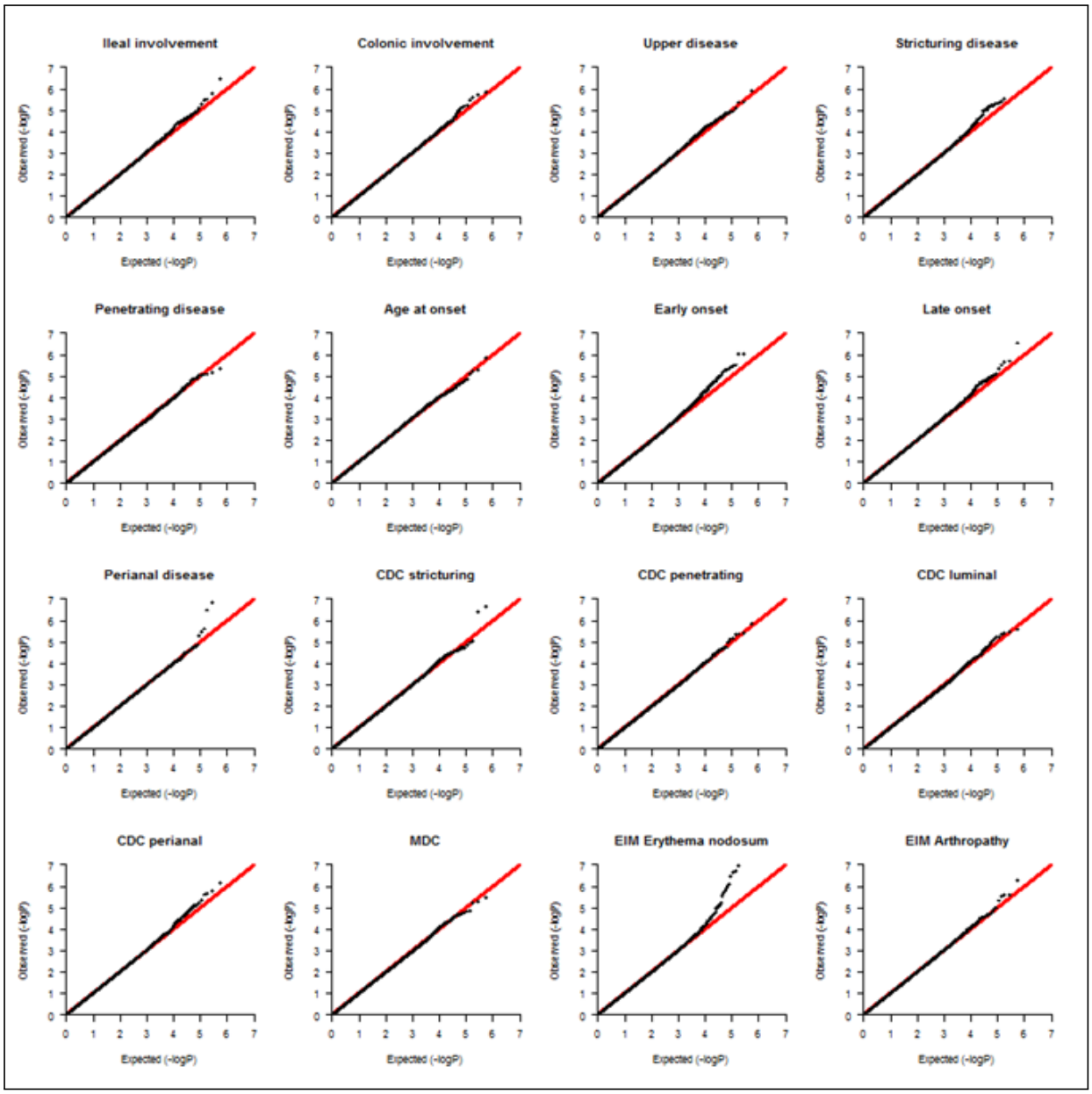


Figure C.7: *Quantile-quantile plots of GWAS analyses of CD phenotypes.* *P*-Values for categorical phenotypes were computed using the χ^2 allelic test. Continuous phenotypes (e.g. age at onset) were evaluated using the Wald test asymptotic *P*-Values.

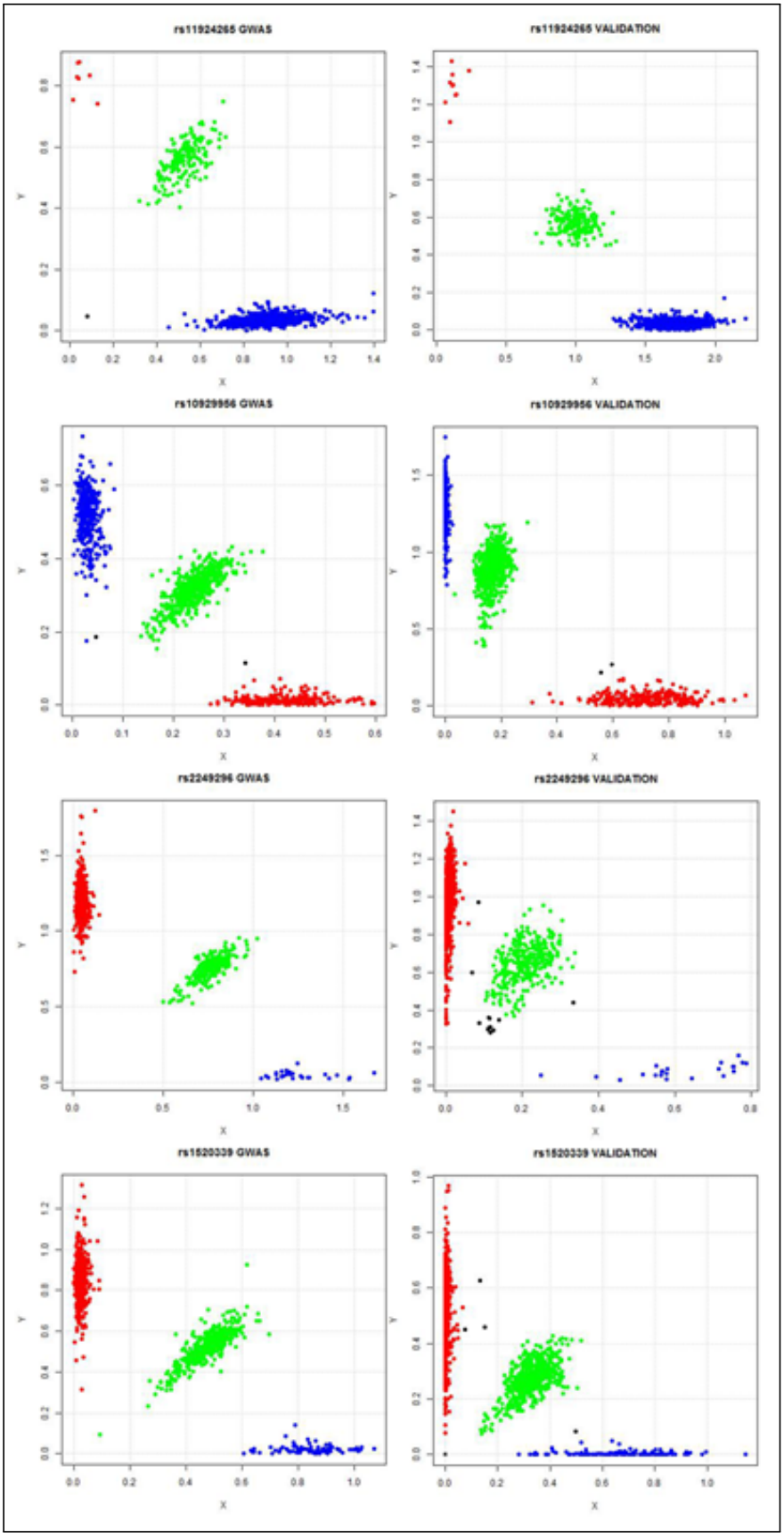


Figure C.8: *Intensity plots for the 4 successfully replicated SNPs.* Channel X and Y intensities for the samples included in the GWAS (i.e. at left) and validation (i.e. at right) cohorts. Samples in black correspond to missing calls while blue, green and red refers to the AA, AB and BB genotypes.

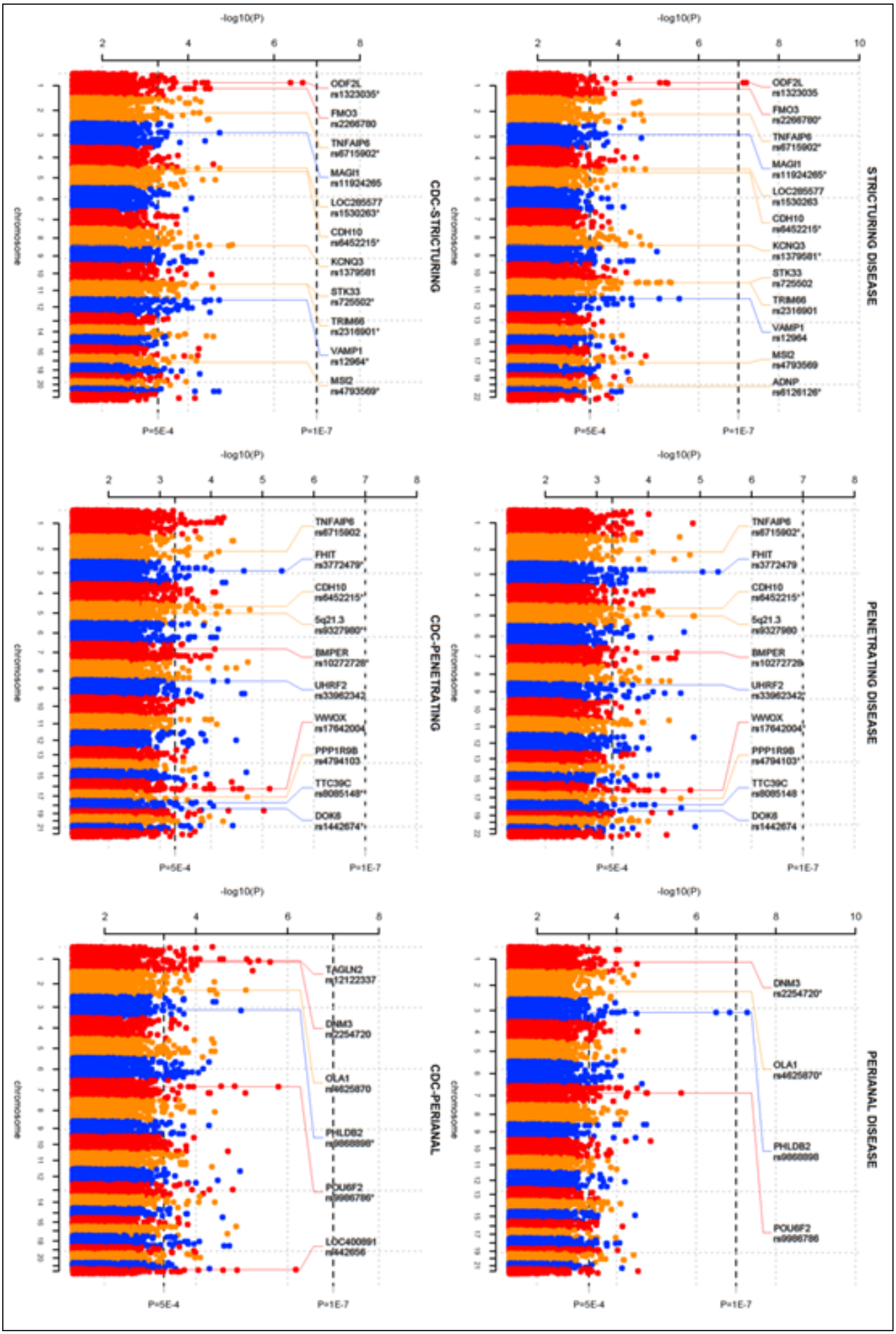


Figure C.9: *Manhattan plots of 6 of the 17 CD phenotypes analyzed.* Manhattan plots of stricturing disease, penetrating disease, perianal disease, CDC-Stricturing, CDC-Penetrating and CDC-Perianal phenotypes GWASs. For each phenotype, SNPs selected for replication are annotated; those with an asterisk at the end were selected due to its association with another phenotype of study but also obtained moderate association values for the analyzed phenotype. Dashed black lines represent $P = 5 \cdot 10^{-4}$ and $P = 1 \cdot 10^{-7}$ levels of signification.

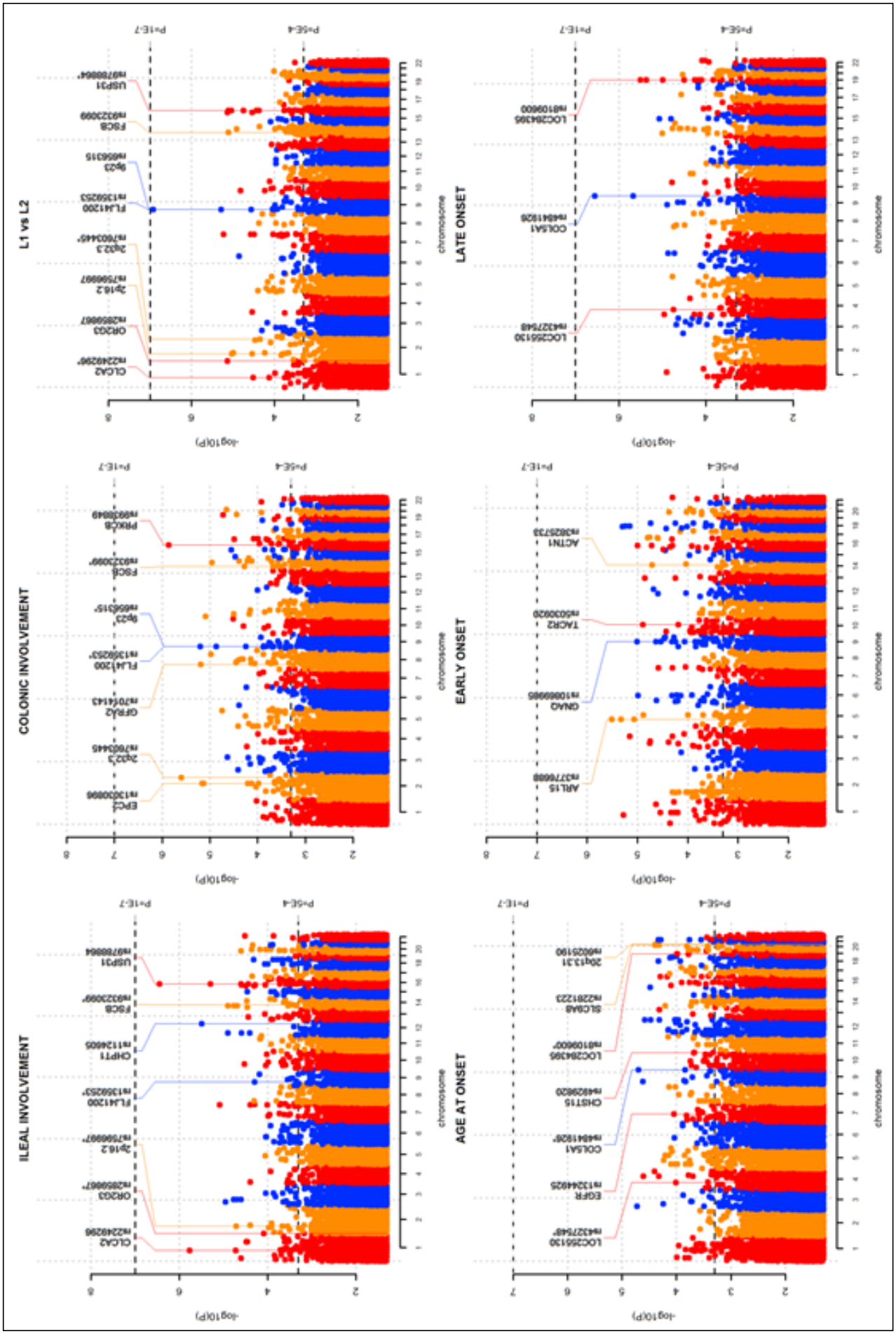


Figure C.10: *Manhattan plots of 6 of the 17 CD phenotypes analyzed.* Manhattan plots of ileal involvement, colonic involvement, purely ileal vs purely colonic location, age at onset, early onset and late onset phenotypes GWASs. For each phenotype, SNPs selected for replication are annotated; those with an asterisk at the end were selected due to its association with another phenotype of study but also obtained moderate association values for the analyzed phenotype. Dashed black lines represent $P = 5 \cdot 10^{-4}$ and $P = 1 \cdot 10^{-7}$.

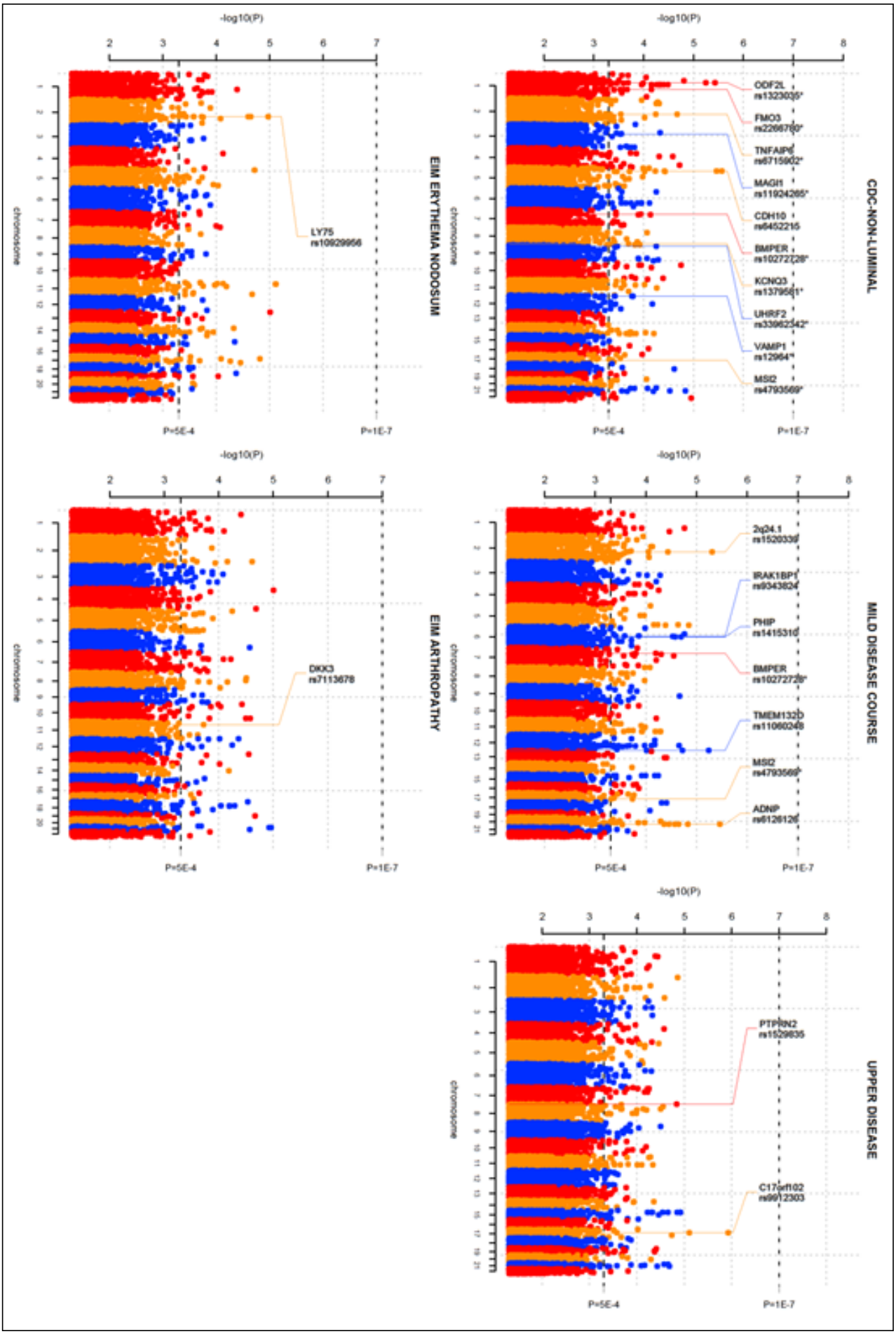


Figure C.11: *Manhattan plots of 5 of the 17 CD phenotypes analyzed.* Manhattan plots of CDC-Non-Luminal, mild disease course, upper disease, erythema nodosum and peripheral arthropathy phenotypes GWASs. For each phenotype, SNPs selected for replication are annotated; those with an asterisk at the end were selected due to its association with another phenotype of study but also obtained moderate association values for the analyzed phenotype. Dashed black lines represent $P = 5 \cdot 10^{-4}$ and $P = 1 \cdot 10^{-7}$ levels of signification.

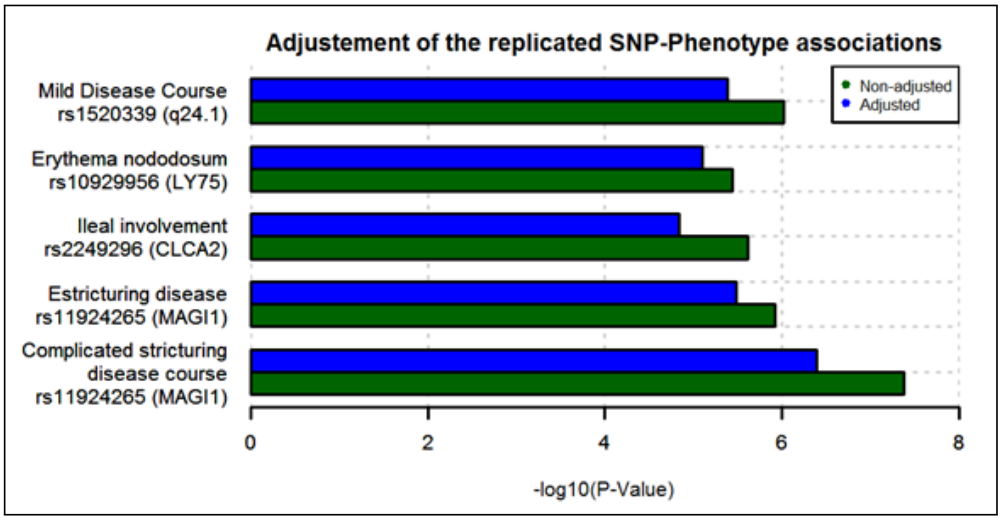


Figure C.12: *Adjusted association of replicated SNP-Phenotype associations.* Unadjusted and adjusted *P*-Values for the four SNP-Phenotype validated associations.

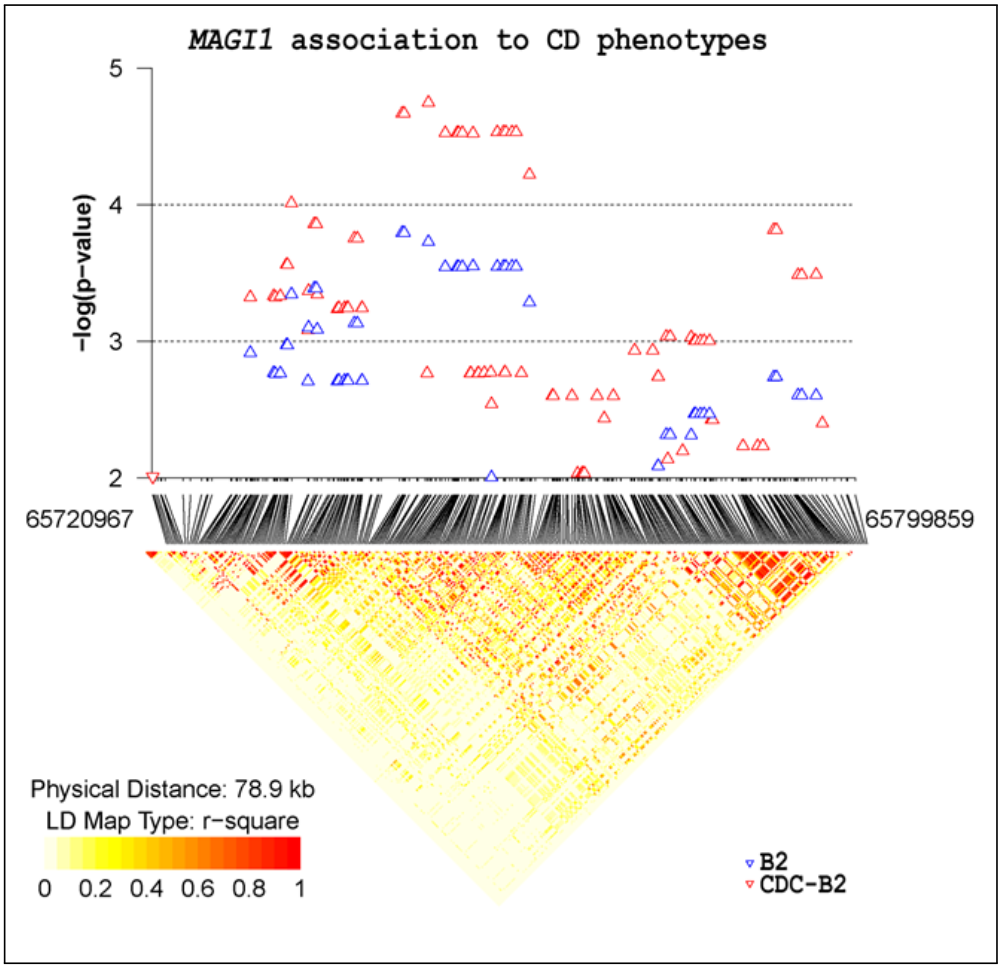


Figure C.13: *GWAS results within the imputed MAGI1 locus.* GWAS association values with structuring behaviour (B2), complicated stricturing disease course (CDC-B2), and LD pattern within the *MAGI1* locus. The upward and downward triangles indicate that the minor allele is a risk or a protective allele, respectively.

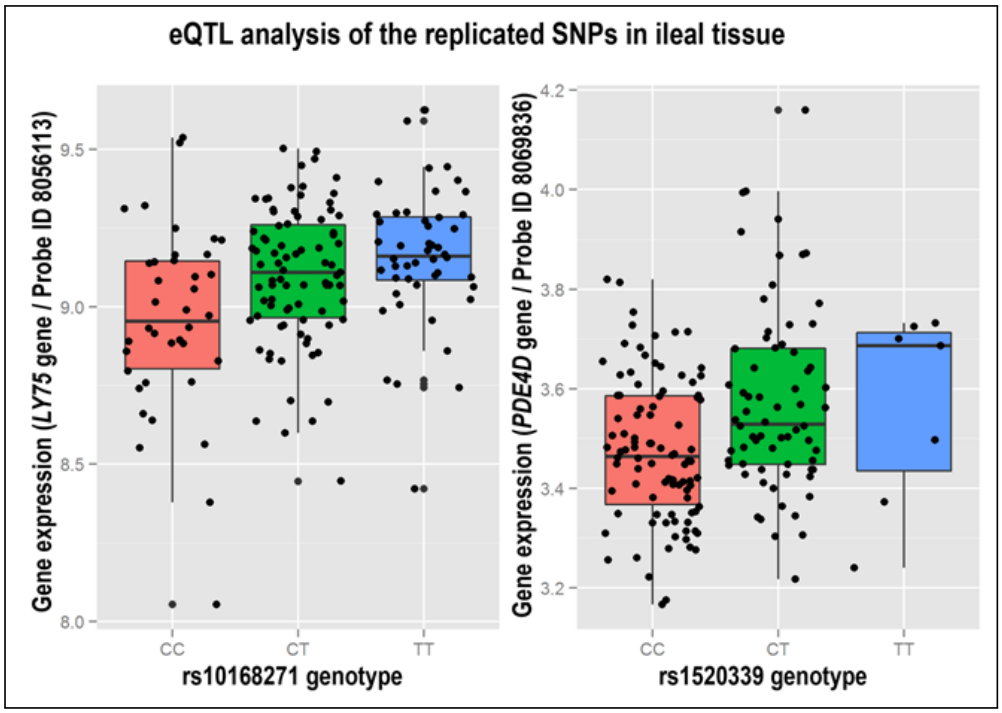


Figure C.14: *eQTL analysis of the replicated phenotype associated SNPs.* This figure shows two relevant ileal eQTLs found for two of the associated SNPs. The NCBI GEO GSE41269 dataset was used for this analysis.

C.2 Supplementary tables

Reference	Complete citation
Cleynen (2013)	Cleynen I, Gonzalez JR, Figueroa C, et al. **Genetic factors conferring an increased susceptibility to develop Crohn's disease also influence disease phenotype: results from the IBDchip European Project.** Gut. 2013;62:1556-65.
Cummings (2010)	Cummings JRF, Cooney RM, Clarke G, et al. **The genetics of NOD-like receptors in Crohn's disease.** Tissue Antigens 2010;76:48-56.1
Cuthbert (2002)	Cuthbert AP, et al. **The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease.** Gastroenterol. 2002;122:867-874.
Dambacher (2007)	Dambacher J, Staudinger T, Seiderer J, et al. **Macrophage migration inhibitory factor (MIF) -173G/C promoter polymorphism influences upper gastrointestinal tract involvement and disease activity in patients with Crohn's disease.** Inflamm Bowel Dis. 2007;13:71-82.1
Duraes (2012)	Duraes C, Machado JC, Portela F, et al. **Phenotype-genotype profiles in Crohn's disease predicted by genetic markers in autophagy-related genes (GOIA study II).** Inflammatory Bowel Diseases 2013;19:230-9.
Eglinton (2012)	Eglinton TW, Roberts R, Pearson J, et al. **Clinical and Genetic Risk Factors for Perianal Crohn's Disease in a Population-Based Cohort.** Am J Gastroenterol 2012;107:589-596.
Fowler (2014)	Fowler SA, Ananthakrishnan AN, Gardet A, et al. **SMAD3 gene variant is a risk factor for recurrent surgery in patients with Crohn's disease.** Journal of Crohn's and Colitis. 2014;8:845-51
Gazouli (2010)	Gazouli M, Pachoula I, et al. **NOD2/CARD15, ATG16L1 and IL23R gene polymorphisms and childhood-onset of Crohn's disease.** World J Gastroenterol. 2010;16:1753-8.
Glas (2010)	Glas J, Seiderer J, Nagy M, et al. Evidence for STAT4 as a common autoimmune gene: rs7574865 is associated with colonic Crohn's disease and early disease onset. PLoS One. 2010;5:e10373.
Glas (2012)	Glas J, Seiderer J, Wagner J, et al. **Analysis of IL12B Gene Variants in Inflammatory Bowel Disease.** PLoS ONE 2012;7:e34349.
Glas (2012b)	Glas J, Wagner J, Seiderer J, et al. **PTPN2 Gene Variants Are Associated with Susceptibility to Both Crohn's Disease and Ulcerative Colitis Supporting a Common Genetic Disease Background.** PLoS ONE 2012;7:e33682.
Henckaerts (2009)	Henckaerts L, Van Steen K, Verstreken I, et al. **Genetic risk profiling and prediction of disease course in Crohn's disease patients.** Clin Gastroenterol Hepatol. 2009;7:972-980.e2.
Imielinski (2009)	Imielinski M, Baldassano RN, Griffiths A, et al. **Common variants at five new loci associated with early-onset inflammatory bowel disease.** Nat Genet. 2009;41:1335-40.
Jung (2012)	Jung C, Colombel J-F, Lemann M, et al. **Genotype/Phenotype Analyses for 53 Crohn's Disease Associated Genetic Polymorphisms.** PLoS ONE 2012;7:e52223.
Koslowski (2009)	Koslowski MJ, Kubler I, Chamaillard M, et al. **Genetic variants of Wnt transcription factor TCF-4 (TCF7L2) putative promoter region are associated with small intestinal Crohn's disease.** PLoS One. 2009;4:e4496.
Kugathasan (2008)	Kugathasan S, Baldassano RN, Bradfield JP, et al. **Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease.** Nat Genet. 2008;40:1211-5.
Latiano (2009)	Latiano A, Palmieri O, Cucchiara S, et al. **Polymorphism of the IRGM gene might predispose to fistulizing behavior in Crohn's disease.** Am J Gastroenterol. 2009;104:110-6.
Lee (2013)	Lee JC, Espeli M, Anderson CA, et al. **Human SNP Links Differential Outcomes in Inflammatory and Infectious Disease to a FOXO3-Regulated Pathway.** Cell 2013;155:57-69.
Marcil (2012)	Marcil V, Sinnett D, Seidman E, et al. **Association between genetic variants in the HNF4A gene and childhood-onset Crohn's disease.** Genes Immun 2012;13:556-565.
Pierik (2005)	Pierik M, Joossens S, Van Steen K, et al. **Toll-like receptor-1, -2, and -6 polymorphisms influence disease extension in inflammatory bowel diseases.** Inflammatory Bowel Diseases 2006;12:1-8.
Prescott (2007)	Prescott NJ, Fisher SA, Franke A, et al. **A nonsynonymous SNP in ATG16L1 predisposes to ileal Crohn's disease and is independent of CARD15 and IBD5.** Gastroenterology. 2007;132:1665-71.
Ridder (2007)	de Ridder L, Weersma RK, Dijkstra G, et al. **Genetic susceptibility has a more important role in pediatric-onset Crohn's disease than in adult-onset Crohn's disease.** Inflamm Bowel Dis. 2007;13:1083-92.
Sehgal (2012)	Sehgal R, Berg A, Polinski JI, et al. **Mutations in IRGM are associated with more frequent need for surgery in patients with ileocolonic Crohn's disease.** Dis Colon Rectum. 2012;55:115-21.
Simms (2010)	Simms LA, Doecke JD, Roberts RL, et al. **KCNN4 gene variant is associated with ileal Crohn's Disease in the Australian and New Zealand population.** Am J Gastroenterol. 2010;105:2209-17.
Thalmaier (2006)	Thalmaier D, Dambacher J, Seiderer J, et al. **The +1059G/C polymorphism in the C-reactive protein (CRP) gene is associated with involvement of the terminal ileum and decreased serum CRP levels in patients with Crohn's disease.** Aliment Pharmacol Ther. 2006;24:1105-15.
Wang (2011)	Wang AH, Lam W-J, Han D-Y, et al. **The effect of IL-10 genetic variation and interleukin 10 serum levels on Crohn's disease susceptibility in a New Zealand population.** Human Immunology 2011;72:431-435.
Waschke (2005)	Waschke KA, Villani AC, Vermeire S, et al. **Tumor necrosis factor receptor gene polymorphisms in Crohn's disease: association with clinical phenotypes.** Am J Gastroenterol. 2005;100:1126-33.

Table C.1: Previous studies on Crohn's disease subphenotypes.

Phenotype	Previously reported association				Top-associated SNP within the NOD2 locus					Validation at the locus level			
	Variant[1]	P	OR (95%CI)	Reference	SNP	Basepair	P_{min}	OR (95%CI)	P_{minC}[2]	N_{SIG}	N_{ISIG}	$P_{gene-set}$	$SNPs_{ISIG}$
Bowel resection	SNP_C	$2.28e^{-5}$	1.73 (1.35-2.24)	Cleynen (2013)	rs62029864	50752164	$4.45e^{-6}$	1.58 (1.30-1.93)	$5.92e^{-4}$	52	3	$1.68e^{-3}$	rs62029864 rs1981760 rs1861759
Late disease onset	-		-	-	rs13380733	50742681	$1.40e^{-4}$	0.61 (0.47-0.79)	$1.86e^{-2}$	18	3	$2.31e^{-2}$	rs13380741 rs5743291 rs9940175
Purely Ileal vs Purely Colonic location	SNP_C	$4.00e^{-4}$	0.49 (0.30-0.81)	Cuthbert (2002)	rs2066850	50730229	$1.21e^{-4}$	0.54 (0.39-0.74)	$1.61e^{-2}$	13	3	$4.92e^{-3}$	rs2066850 rs8057341 rs5743289
Ileal involvement	SNP_C	$2.02e^{-6}$	1.90 (1.46-2.47)	Cleynen (2013)	rs5743266	50731096	$4.19e^{-4}$	1.65 (1.25-2.18)	$5.57e^{-2}$	21	3	$2.41e^{-2}$	rs5743266 rs8057341 rs5743289
Inflammatory behaviour (B1)	-		-	-	rs5743289	50756774	$2.13e^{-3}$	0.71 (0.57-0.88)	$2.83e^{-1}$	15	3	$6.24e^{-2}$	rs5743289 rs7500036 rs5743259
Colon involvement	SNP_C	$6.96e^{-4}$	0.57 (0.41-0.78)	Cleynen (2013)	rs2076756	50756881	$6.06e^{-3}$	0.73 (0.59-0.92)	$8.07e^{-1}$	23	3	$2.35e^{-1}$	rs2076756 rs2004804 rs2066844
Stricturing behaviour (B2)	SNP_C	$3.16e^{-6}$	1.82 (1.42-2.35)	Cleynen (2013)	rs62027635	50739450	$6.16e^{-3}$	1.36 (1.09-1.69)	$8.19e^{-1}$	17	3	$1.59e^{-1}$	rs62027635 rs5743289 rs13339578
Penetrating behaviour (B3)	SNP_C	$1.26e^{-2}$	1.25 (1.05-1.49)	Cleynen (2013)	rs5743289	50756774	$6.15e^{-3}$	1.42 (1.10-1.84)	$8.18e^{-1}$	12	3	$2.82e^{-1}$	rs5743289 rs5743266 rs5743291
Early disease onset	SNP13	$6.70e^{-3}$	1.87 (1.21-2.88)	Gazouli (2010)	rs141612356	50763510	$1.59e^{-3}$	4.53 (1.62-12.63)	$2.11e^{-1}$	5	3	$1.10e^{-1}$	rs141612356 rs9925315 rs17312836

[1] SNP8 (rs2066844), SNP12 (rs2066845) and SNP13 (rs2066847) are usually condensed in one variable (SNP_C) that accounts for the number of risk alleles in these three SNPs.

[2] Corrected by the 133 SNPs analyzed within the *NOD2* locus

Significant *NOD2* associations are marked in bold.

Table C.2: Subphenotype associations at the *NOD2* locus.

LOCI	SUBPHENOTYPE	SNP	Basepair	P[1]	LD[2]
q24.1	MDC	rs1540803	156709593	$3.69 \cdot 10^{-6}$	0.99
q24.1	MDC	**rs1520339**	156751921	$4.94 \cdot 10^{-6}$	1.00
q24.1	MDC	rs1540804	156725938	$2.64 \cdot 10^{-5}$	0.95
q24.1	MDC	rs3843321	156762304	$2.92 \cdot 10^{-5}$	0.95
q24.1	MDC	rs10190879	156693033	$7.21 \cdot 10^{-5}$	0.93
MAGI1	CDC-B2	rs77617242	65751923	$1.29 \cdot 10^{-5}$	0.98
MAGI1	CDC-B2	rs114897644	65755145	$2.07 \cdot 10^{-5}$	1.00
MAGI1	CDC-B2	rs79650565	65755317	$2.07 \cdot 10^{-5}$	1.00
MAGI1	CDC-B2	rs77474762	65755741	$2.07 \cdot 10^{-5}$	1.00
MAGI1	CDC-B2	chr3:65759650:D	65759650	$2.07 \cdot 10^{-5}$	1.00
MAGI1	CDC-B2	**rs11924265**	65760342	$2.07 \cdot 10^{-5}$	1.00
MAGI1	CDC-B2	rs11920326	65760606	$2.07 \cdot 10^{-5}$	1.00
MAGI1	CDC-B2	rs79032699	65761301	$2.07 \cdot 10^{-5}$	1.00
MAGI1	CDC-B2	rs72908267	65761770	$2.07 \cdot 10^{-5}$	1.00
MAGI1	CDC-B2	rs79225788	65753811	$2.07 \cdot 10^{-5}$	0.98
LY75	Erythema nod.	rs12692570	160726599	$5.81 \cdot 10^{-6}$	0.97
LY75	Erythema nod.	chr2:160726952:I	160726952	$6.19 \cdot 10^{-6}$	0.97
LY75	Erythema nod.	chr2:160726954:I	160726954	$6.19 \cdot 10^{-6}$	0.97
LY75	Erythema nod.	rs2136977	160722850	$8.54 \cdot 10^{-6}$	0.99
LY75	Erythema nod.	rs2729704	160707866	$9.53 \cdot 10^{-6}$	0.99
LY75	Erythema nod.	rs10168271	160729186	$9.69 \cdot 10^{-6}$	0.99
LY75	Erythema nod.	**rs10929956**	160714615	$1.05 \cdot 10^{-5}$	1.00
LY75	Erythema nod.	rs55664214	160708059	$1.06 \cdot 10^{-5}$	1.00
CLCA2	Ileal involv.	rs11161820	86900760	$9.52 \cdot 10^{-7}$	0.95
CLCA2	Ileal involv.	rs1334150	86897973	$1.18 \cdot 10^{-6}$	0.92
CLCA2	Ileal involv.	**rs2249296**	86910264	$1.71 \cdot 10^{-6}$	1.00
CLCA2	Ileal involv.	rs2252313	86908149	$1.82 \cdot 10^{-6}$	1.00
CLCA2	Ileal involv.	rs9433030	86901465	$3.14 \cdot 10^{-6}$	0.95
CLCA2	Ileal involv.	rs2791464	86911543	$4.08 \cdot 10^{-6}$	1.00
CLCA2	Ileal involv.	rs4292929	86896354	$5.27 \cdot 10^{-6}$	0.91

[1] Association *P*-Values on imputed SNPs.
[2] LD values (r^2) with the originally associated SNP in the matched control population.
Bolded SNPs are the originally associated SNPs.

Table C.3: Association *P*-Values and LD for the imputed SNPs in the replicated loci.

SNP	Phenotype	Probe ID	Position	Gene	P	β
rs10168271	Erythema nodosum	7940580	chr11:61705366-61705469	-	$3.45 \cdot 10^{-05}$	-0.073
rs10168271	Erythema nodosum	8056113	chr2:160659872-160761247	LY75	$1.70 \cdot 10^{-04}$	-0.099
rs10168271	Erythema nodosum	8100990	chr4:74919755-74921116	PPBPL2	$2.33 \cdot 10^{-04}$	-0.068
rs10168271	Erythema nodosum	8060255	chr2:242304241-242304314	-	$6.31 \cdot 10^{-04}$	0.075
rs10168271	Erythema nodosum	8075479	chr22:31618722-31618839	-	$7.26 \cdot 10^{-04}$	0.084
rs1520339	Mild disease course	7939867	chr11:48266646-48267567	OR4X2	$9.93 \cdot 10^{-06}$	-0.12
rs1520339	Mild disease course	8067844	chr21:14982175-15013906	POTED	$1.62 \cdot 10^{-05}$	-0.076
rs1520339	Mild disease course	8058614	chr2:211295975-211341431	LANCL1	$8.76 \cdot 10^{-05}$	0.085
rs1520339	Mild disease course	8069991	chr21:33948862-33957843	TCP10L	$9.86 \cdot 10^{-05}$	-0.083
rs1520339	Mild disease course	7919208	chr1:145509235-145516076	GNRHR2	$1.17 \cdot 10^{-04}$	-0.083
rs1520339	Mild disease course	8112220	chr5:58266711-59783890	PDE4D	$1.38 \cdot 10^{-04}$	0.084
rs1520339	Mild disease course	8069836	chr21:31869142-31869472	KRTAP19-4	$1.46 \cdot 10^{-04}$	-0.099
rs1520339	Mild disease course	8115375	chr5:153371263-153418497	FAM114A2	$2.48 \cdot 10^{-04}$	0.096
rs1520339	Mild disease course	8069811	chr21:31720717-31720924	KRTAP23-1	$2.53 \cdot 10^{-04}$	-0.076
rs1520339	Mild disease course	7945956	chr11:4308729-4309175	LOC143506	$2.56 \cdot 10^{-04}$	-0.091
rs1520339	Mild disease course	7918757	chr1:115090582-115090818	DENND2C	$3.62 \cdot 10^{-04}$	-0.114
rs1520339	Mild disease course	8176910	chrY:27601452-27606322	GOLGA2P2Y	$3.97 \cdot 10^{-04}$	-0.058
rs1520339	Mild disease course	8177413	chrY:26356114-26360984	GOLGA2P2Y	$3.97 \cdot 10^{-04}$	-0.058
rs1520339	Mild disease course	7918157	chr1:108113782-108507585	VAV3	$4.00 \cdot 10^{-04}$	0.13
rs1520339	Mild disease course	7922756	chr1:183217372-183387737	NMNAT2	$4.24 \cdot 10^{-04}$	-0.082
rs1520339	Mild disease course	8178582	chr6:31680113-31681589	LY6G6E	$4.42 \cdot 10^{-04}$	-0.077
rs1520339	Mild disease course	8179810	chr6:31680113-31681589	LY6G6E	$4.42 \cdot 10^{-04}$	-0.077
rs1520339	Mild disease course	7945954	chr11:4263286-4263870	LOC143506	$4.55 \cdot 10^{-04}$	-0.078
rs1520339	Mild disease course	8075059	chr22:26360023-26360239	-	$6.07 \cdot 10^{-04}$	-0.104
rs1520339	Mild disease course	8064168	chr20:62367995-62370456	LIME1	$7.41 \cdot 10^{-04}$	-0.064
rs1520339	Mild disease course	8113073	chr5:90664541-90679121	ARRDC3	$7.60 \cdot 10^{-04}$	0.134
rs1520339	Mild disease course	7932786	chr10:28696968-28697180	-	$8.43 \cdot 10^{-04}$	-0.096
rs1520339	Mild disease course	8177083	chrY:8240142-8240212	-	$9.03 \cdot 10^{-04}$	-0.125
rs1520339	Mild disease course	8039625	chr19:57645459-57656570	ZIM3	$9.32 \cdot 10^{-04}$	-0.085
rs2249296	Ileal involvement	7929541	chr10:97759883-97792441	CC2D2B	$1.21 \cdot 10^{-04}$	0.198
rs2249296	Ileal involvement	8106098	chr5:71403118-71502853	MAP1B	$1.38 \cdot 10^{-04}$	0.152
rs2249296	Ileal involvement	8108683	chr5:140474219-140476964	PCDHB2	$5.13 \cdot 10^{-04}$	0.133
rs2249296	Ileal involvement	8172914	chrX:53559057-53713673	HUWE1	$5.54 \cdot 10^{-04}$	0.069
rs2249296	Ileal involvement	8068520	chr21:38417831-38417933	-	$6.10 \cdot 10^{-04}$	-0.114
rs2249296	Ileal involvement	8080100	chr3:51661676-51697634	RAD54L2	$6.54 \cdot 10^{-04}$	0.081
rs2249296	Ileal involvement	8126943	chr6:50011288-50016364	DEFB112	$7.70 \cdot 10^{-04}$	-0.104
rs2249296	Ileal involvement	8098289	chr4:169926400-169926493	-	$9.78 \cdot 10^{-04}$	-0.225

Table C.4: eQTL analysis of the replicated SNPs in ileal tissue (NCBI GEO GSE40929)

C.3 Supplementary material and methods

C.3.1 Univariate and multivariate analysis of CD phenotypic co-occurrence

Association analyses between the different CD phenotypes were performed on the cohorts used for the GWAS and Replication analyses. The analyses were done using univariate and multivariate logistic regression methods. In addition to the defined CD phenotypes (Tables 5.1 and 5.2), epidemiological variables such as gender, familial CD (i.e. $\geq$1 first- or second-degree relative diagnosed with CD) and smoking habit (i.e. smokers at the time of diagnosis) were also included in these analyses. The following Table C.5 shows the studied phenotypes (i.e. dependent variables) and the set of variables included as independent variables (i.e. risk factors or potential confounders) in the univariate and multivariate analyses.

Phenotype	Variables included in the univariate analysis	Variables included in the multivariate analysis
Stricturing disease (B2)	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Ileal involvement; Colonic involvement; Ileocolonic; Upper disease	Sex; Age at onset; Smoking habit; Ileal involvement; Perianal disease; Familial CD
Penetrating disease (B3)	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Ileal involvement; Colonic involvement; Ileocolonic; Upper disease	Sex; Age at onset; Smoking habit; Ileal involvement; Perianal disease; Familial CD
Complicated stricturing disease (CDC-B2)	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Ileal involvement; Colonic involvement; Ileocolonic; Upper disease	Sex; Age at onset; Smoking habit; Ileal involvement; Perianal disease; Familial CD
Complicated penetrating disease (CDC-B3)	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Ileal involvement; Colonic involvement; Ileocolonic; Upper disease	Sex; Age at onset; Smoking habit; Ileal involvement; Perianal disease; Familial CD
Upper disease (L4)	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Ileal involvement; Colonic involvement; Ileocolonic	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Ileal involvement
Ileal involvement	Sex; Smoking habit; Familial CD; Age at onset	Sex; Smoking habit; Familial CD; Age at onset
Colonic involvement	Sex; Smoking habit; Familial CD; Age at onset	Sex; Smoking habit; Familial CD; Age at onset
Ileocolonic	Sex; Smoking habit; Familial CD; Age at onset	Sex; Smoking habit; Familial CD; Age at onset
Perianal disease	Sex; Smoking habit; Familial CD; Age at onset; B2; B3; Ileal involvement; Colonic involvement; Ileocolonic; Upper disease	Sex; Smoking habit; Familial CD; Age at onset; B2; Colonic involvement
Bowel resection	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Stricturing or penetrating disease; Ileal involvement; Colonic involvement; Ileocolonic	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Stricturing or penetrating disease; Ileal involvement
Erythema nodosum	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Stricturing or penetrating disease; Ileal involvement; Colonic involvement; Ileocolonic; Upper disease	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Stricturing or penetrating disease; Ileal involvement; Upper disease
Arthropathy	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Stricturing or penetrating disease; Ileal involvement; Colonic involvement; Ileocolonic; Upper disease	Sex; Smoking habit; Perianal disease; Familial CD; Age at onset; Stricturing or penetrating disease; Ileal involvement; Upper disease

Table C.5: Variables included as independent variables in the univariate and multivariate analyses.

First, the association between each phenotype of interest and its corresponding risk factors was studied in separate univariate models. The strength of the association in these analyses was used for selecting the variables to be included in the multivariate analysis, avoiding potential colinearity due to correlated phenotypes. For example, disease behaviour phenotypes (B2/B3/CDC-B2/CDC-B3) showed a strong association to ileal involvement. Consequently, colonic involvement, ileocolonic location and upper disease were excluded from the multivariate analysis due to their strong correlation with the ileal involvement phenotype.

Subsequently, a multivariate logistic regression model was fitted for each phenotype accounting for the variables previously selected from the univariate analysis.

The resulting *P*-Values and ORs for each multivariate analysis in the combined cohort are given in Supplementary Table 1. Supplementary Figures C.2 to C.4 provide all the ORs and significant associations in the univariate and multivariate analyses for each analyzed cohort (GWAS, Replication and Combined cohorts).

C.3.2 Genotyping in Discovery and Replication Analysis

Genome-wide genotyping was performed using the Illumina Quad610 Beadchips (Illumina, San Diego, CA, USA) at the Centro Nacional de Genotipado (CeGen, Spain). Single nucleotide polymorphism genotype calling was performed using the Illumina GenomeStudio software v2010.1 (Illumina, San Diego, CA, USA). After excluding mitochondrial, X and Y chromosome markers or markers not mapped on the latest human assembly GRCh37/hg19 (i.e. dismissed SNPs due to changes in the human reference genomic sequences), a total of 576,818 markers from the 616,794 available markers were selected for quality control (QC) analysis. Only SNPs that had a >95% call rate (98.1%), a minor allele frequency (MAF) >0.01 (95.8%), and that showed Hardy-Weinberg equilibrium (99.7% of SNP, $P > 0.0001$ in healthy controls) were considered for further analysis. All the samples included in the initial cohort (n=1,338) showed no outlier heterozygosity rates and sex concordance between database records and X chromosome heterozygosity. We established an inclusion threshold of >95% call rate (n=1,330, 99.4% of samples passed). To evaluate potential population stratification we performed a principal component analysis based on the EIGENSTRAT method[214] and outlier samples on this basis (>3 standard deviations above the mean on the first or second principal component) were excluded (95.8% of samples passed; Supplementary Figure C.1). Additionally, 184 CD samples were excluded due to a high missingness on their clinical data records. After QC analysis, a final dataset of 539,846 SNPs and 1,090 CD patients was available for the clinical subphenotype GWAS.

Genotype imputation was used in this study to: (a) expand the number of analyzed SNPs at the *NOD2* locus, (b) identify neighboring imputed SNPs that reach higher levels of association than the formerly associated microarray SNPs in the GWAS stage, and, (c) to better

estimate the limits of the genomic associated regions around each associated SNP that has been replicated. Imputation was performed in the GWAS cohort using SHAPEIT V2-644[215] and IMPUTE V2[216] softwares. The reference genetic data for imputation was obtained from the 1000 Genomes Project[217]. We specifically used the data of the latest release (phase 1, version 3) from the European cohort (i.e. 379 samples). QC measures for imputed SNPs included MAF >0.05 and a corresponding IMPUTE2 info quality metric was set to >0.8.

Fifty-seven SNPs were selected for replication and genotyped using the Illumina GoldenGate assay (Illumina, USA). From these, nine SNPs did not exceed the minimum assay quality score threshold. In these cases, neighboring SNPs in high linkage disequilibrium (LD) with the original SNP (r^2>0.9 on Caucasian European Hapmap samples from the 1000 Genomes Project) and with a high quality score were finally selected. In other five cases, the originally associated SNP was replaced by an imputed SNP at less than 50 kb that exceeded the significance level of the originally associated SNP (≥10-fold *P*-value increment). All the SNPs selected for replication had a call rate over 97.5% (average call rate was 99.61%). From the 1,627 samples, n=225 were excluded due to missing clinical data and n=106 due to low call rates (<90%). The included samples had a missing rate below 10%, from which 90.5% had 0%. Genotyping error rate was estimated using a subset of samples (i.e. 5% of the total) that were genotyped twice. The concordance results between duplicates indicated a genotyping error rate of 0.32%.

C.3.3 Criteria for selection of SNPs for replication

The selection criteria for the 57 SNPs for replication were established according to the statistical significance of the phenotype association, the biological context of the associated SNP and the clinical relevance of the associated phenotype. This task was performed by a multidisciplinary team of clinical and bioinformatician experts using in-house and publicly available data. First, the most statistically significant SNPs ($P < 5 \cdot 10^{-6}$) were directly selected (n=26 SNPs, *P*-Values from $5.16 \cdot 10^{-8}$ to $4.94 \cdot 10^{-6}$). Next, the SNP-Phenotype associations showing a moderate level of significance ($5 \cdot 10^{-6} < P < 5 \cdot 10^{-4}$) were biologically annotated based on:

1. Functional class of the associated SNP as described in the dbSNP database[262] (i.e. intronic, UTR, non-synonymous...).

2. eQTLs identified for the associated SNP, reported by previous studies or in eQTL databases[229;263–268].

3. Overlap with potential regulatory regions based on the ENCODE project data[269].

4. SNPs close to genes (<100Kb) associated with Crohn's disease or related biological processes like inflammation. This was performed by querying PubMed database[270] for articles matching "gene [gene name]" and key words such as "Crohn's disease", "inflammatory bowel disease", "inflammation", "immune system", "colon" and "ileum". The returned results were manually revised.

5. SNPs close to genes (<100Kb) found to be differentially expressed in inflammatory bowel diseases in previous studies[271;272].

6. SNPs close to other SNPs (<100Kb) previously associated to inflammatory bowel diseases and other auto-immune diseases as defined in the NHGRI GWAS Catalog[273].

Using this annotation data, a total of n=31 additional SNPs were selected for replication. The phenotypes of highest clinical interest (i.e. disease course, behaviour or location) were prioritized against other less relevant or infrequent CD phenotypes (i.e. erythema nodosum and peripheral arthropathy). The table C.6 shows the number of SNPs selected for replication according to the phenotype group.

Phenotype group	$N_{selected}$
Age-related phenotypes	11
Behaviour-related phenotypes (including perianal disease)	13
Disease course phenotypes	17
Location phenotypes	14
Extra-intestinal manifestations	2
Total	57

Table C.6: Number of SNPs selected for replication.

Due to the high correlation between some of the phenotypes studied (i.e. ileal involvement and colonic involvement) some of the associated loci showed moderate significance values for multiple phenotypes. In these cases, only the SNP showing the strongest association significance within a locus for any of the phenotypes studied was selected. We previously verified that SNPs in the same locus showing moderate association values were in high LD (r^2>0.8) with it.

C.3.4 Validation of previously reported associations

We performed a validation study of loci previously associated with clinically relevant phenotypes in CD. Excluding *NOD2*, a total 91 SNP-phenotype associations ($P < 0.05$) were identified in multiple studies (Supplementary Table 2). With the exception of the study reporting the *FOXO3* association to mild disease course which evaluated 1,134 SNPs within 81 genes related with IL-2/IL-7 signaling pathways[14], all these studies analyzed the association

of established CD risk loci with the clinical phenotypes. Importantly, most of these genetic associations have still not been replicated in an independent cohort of patients.

In order to evaluate these phenotype associations in our GWAS discovery cohort we performed three validation analyses:

- First, we performed a validation study of the previously associated SNPs with the clinical phenotype of interest. When the previously reported SNP was not available in our curated microarray data, we proceeded to identify a tagSNP to evaluate the association. tagSNPs were identified by evaluating the LD between the available SNPs in our dataset and the previously reported missing SNP. LD was computed using the 1KGP[217] data of CEU samples and the SNP with higher correlation ($r^2 > 0.85$, at least) with the reported SNP was selected as tagSNP.

- Second, a validation analysis exploring the association at the locus level was performed. This analysis takes into account that associations at the SNP level can be easily weakened due to differences in LD patterns when trying to replicate the association on a different population cohort. This analysis was performed using a gene-set based test as implemented in PLINK software[218] (i.e. set-based test). This approach maximizes the statistical power of the analysis (i.e. it takes into account the LD patterns between SNPs) and provides accurate results since it is based on a permutation procedure[274]. For each reported association, we created the SNP set including all the SNPs within the locus (i.e. 50Kb around the reported SNP) and performed the analysis using the standard method parameterization[274] ($P < 0.05$, $N_{SNPs} = 3$, $r^2 = 0.5$ and $N_{perm} = 1000$). The results (Table 5.3 and Supplementary Table 3) provide the number of analyzed SNPs, the number of associated SNPs, the number of independently associated SNPs and the empirical locus association *P*-Value.

- Finally, we also evaluated if the top-ranked association of all the SNPs within the locus was significant after Bonferroni correction (i.e. taking into account the number of SNPs analyzed within each locus).

NOD2 locus is the only one that has been robustly associated to multiple CD phenotypes in previous studies[16] and can therefore be also used as an internal methodological validation. In order to deeply analyze this locus we imputed all the variants within the locus (chr16:50700642-50788313). After the QC of the imputed data a final set of 133 SNPs was ready for analysis. The number of risk variants at rs2066844 (SNP8/R702W), rs2066845 (SNP12/G908R) and rs2066847 (SNP13/1007fs) has been the genetic variant classically associated to CD and its related subphenotypes. However, these SNPs have very low MAFs (i.e.

0.05, 0.01 and 0.01 in the 1KGP European population) and its imputation did not withstand the quality imputation filters. Consequently, we performed the aforementioned second (i.e. gene-set based test) and third (i.e. Bonferroni correction) validation approaches.

C.3.5 eQTL analysis of replicated associated SNPs in ileal tissue

Specific eQTLs on ileal tissue could be investigated using the gene expression and genotyping data available in the GEO NCBI database under the accession number GSE41269[221]. In this study, total RNA was extracted from 173 endoscopically and histologically normal ileal biopsies in patients that underwent ileal pouch-anal anastomosis following colectomy. The included patients consisted of subjects diagnosed with ulcerative colitis or familial adenomatous polyposis. Gene expression data was measured using the Affymetrix Human Gene 1.0 ST Array. This microarray provides the expression levels of 19,047 unique autosomal genes listed in the NCBI database.

In our study, we restricted the eQTL analysis to the four validated loci showing CD phenotype associations. The genotyping data for rs1520339 (i.e. associated to mild disease course) and rs2249296 (i.e. associated to ileal involvement) was available in the GEO dataset. A tagSNP (rs10929956) was also identified for rs10168271 (i.e. associated to Erythema Nodosum) showing a high LD in the European 1KGP population ($r^2 = 0.99$). However, no tagSNPs ($r^2 > 0.8$) were found for rs11924265 (i.e. associated to complicated stricturing disease course) so eQTLs regarding to this SNP could not be evaluated. Table C.7 shows the SNPs used for eQTL evaluation.

Phenotype	SNP	Locus	SNP_{eQTL}	dd (n)	Dd (n)	DD (n)	Model
CDC stricturing	rs11924265	MAGI1	Not found	-	-	-	-
EIM Erythema nodosum	rs10929956	LY75	rs10168271[1]	38	83	48	Genotypic
Ileal involvement	rs2249296	CLCA2	rs2249296	2	38	129	Recessive
MDC	rs1520339	-	rs1520339	7	68	94	Recessive

[1] tagSNP

Table C.7: SNPs used for eQTL evaluation.

The eQTL analysis was performed fitting a linear regression model within the n=169 available samples and the n=13,785 expression probes that mapped to the human genome reference (i.e. as provided by the GEO platform annotation file GPL6244). The identified eQTLs ($P - Value < 1 \cdot 10^{-3}$) are given in Supplementary Table 7.

D | Metabolomics Review Articles

In the following sections we provide the original review articles of metabolomics:

- Section D.1: *Analytical methods in untargeted metabolomics: state of the art in 2015*.
- Section D.2: *Metabolomics in rheumatic diseases*.

D.1 Analytical methods in untargeted metabolomics: state of the art in 2015

frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

REVIEW ARTICLE
published: 05 March 2015
doi: 10.3389/fbioe.2015.00023

Analytical methods in untargeted metabolomics: state of the art in 2015

Arnald Alonso[1,2], Sara Marsal[1] and Antonio Julià[1]*

[1] *Rheumatology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain*
[2] *Department of Automatic Control (ESAII), Polytechnic University of Catalonia, Barcelona, Spain*

Edited by:
Adam James Carroll, The Australian National University, Australia

Reviewed by:
Masahiro Sugimoto, Kei University, Japan
Jianguo Xia, University of British Columbia, Canada

****Correspondence:***
Antonio Julià, Rheumatology Research Group, Vall d'Hebron Research Institute, Baldiri i Reixac, 15-21, Barcelona 08028, Spain
e-mail: toni.julia@vhir.org

Metabolomics comprises the methods and techniques that are used to measure the small molecule composition of biofluids and tissues, and is actually one of the most rapidly evolving research fields. The determination of the metabolomic profile – the metabolome – has multiple applications in many biological sciences, including the developing of new diagnostic tools in medicine. Recent technological advances in nuclear magnetic resonance and mass spectrometry are significantly improving our capacity to obtain more data from each biological sample. Consequently, there is a need for fast and accurate statistical and bioinformatic tools that can deal with the complexity and volume of the data generated in metabolomic studies. In this review, we provide an update of the most commonly used analytical methods in metabolomics, starting from raw data processing and ending with pathway analysis and biomarker identification. Finally, the integration of metabolomic profiles with molecular data from other high-throughput biotechnologies is also reviewed.

Keywords: metabolomics, nuclear magnetic resonance, mass spectrometry, untargeted, spectral processing, data analysis, pathway analysis, integration

INTRODUCTION

Metabolomics is the study of the metabolite composition of a cell type, tissue, or biological fluid. The analysis of the complete set of metabolites – the metabolome – has been present in biological research for more than a decade (Patti et al., 2012). However, major recent advances in the technologies used to extract and analyze this type of molecular data have revolutionized its applicability in the analysis of organisms and relevant biological processes (Zhang et al., 2012). To date, metabolomics is envisaged as one of the major "omics" tools that will most contribute into challenging research objectives like the personalization of treatments in medical practice.

The metabolites are the intermediates or end products of multiple enzymatic reactions and therefore are the most informative proxies of the biochemical activity of an organism. The present technologies are allowing the study of tens to hundreds of metabolites in complex biological samples (Patti et al., 2012). One of the facts that is most contributing to the rapid growth of metabolomics is its wide range of applications. These applications cover diverse research areas like plant biology (Qi and Zhang, 2014), nutrition (Orešič, 2009; Gibbons et al., 2015), animal breeding (Kühn, 2012), drug discovery (Robertson and Frevert, 2013; Kell and Goodacre, 2014), and the study of human diseases (Kaddurah-Daouk et al., 2008; Mamas et al., 2011). The biomedical field is actually one of the most active areas of development in metabolomics, and includes the search for diagnostic and prognostic biomarkers as well as predictors of treatment response (Meyer et al., 2013; Armitage and Barbas, 2014; Julià et al., 2014). Also in this field, the use of metabolomics is helping to characterize the impact of key environmental factors on human health. In this area, one of the most promising applications is the characterization of gut–microbiota interactions in humans (Wikoff et al., 2009; Nicholson et al., 2012).

To date, the two main technical approaches for the generation of metabolomic data are nuclear magnetic resonance (NMR) and mass spectrometry (MS; Fuhrer and Zamboni, 2015). NMR is a fast and highly reproducible spectroscopic technique that is based on the energy absorption and re-emission of the atom nuclei due to variations in an external magnetic field (Bothwell and Griffin, 2011). Depending on the atom nuclei being targeted by the applied magnetic field, different types of metabolomic data are generated. However, in the analysis of samples of biological origin, hydrogen is the most commonly targeted nucleus (^{1}H-NMR), due to its natural abundance in biological samples. Although less frequent, other atoms like carbon (^{13}C-NMR) and phosphorus (^{31}P NMR) are also targeted by NMR, providing additional information on specific metabolite types (Reo, 2002).

The resulting spectral data in NMR not only allows the quantification of the concentration of metabolites but also provides information about its chemical structure. The spectral peak areas generated by each molecule are used as an indirect measure of the quantity of the metabolite in the sample, while the pattern of spectral peaks informing on the physical properties of the molecule is used to identify the type of metabolite. The spectral data obtained with NMR techniques can be referenced to one or two frequency axes. One dimensional NMR (1D-NMR) spectra are based on a single frequency axis, where the peaks of each molecule are placed within its resonant frequencies (**Figure 1**). 1D-NMR is the most commonly used method in high-throughput metabolomics studies. Conversely, two dimensional NMR (2D-NMR) spectra are based on two frequency axis, and its use is often restricted to the characterization of those compounds that cannot be identified

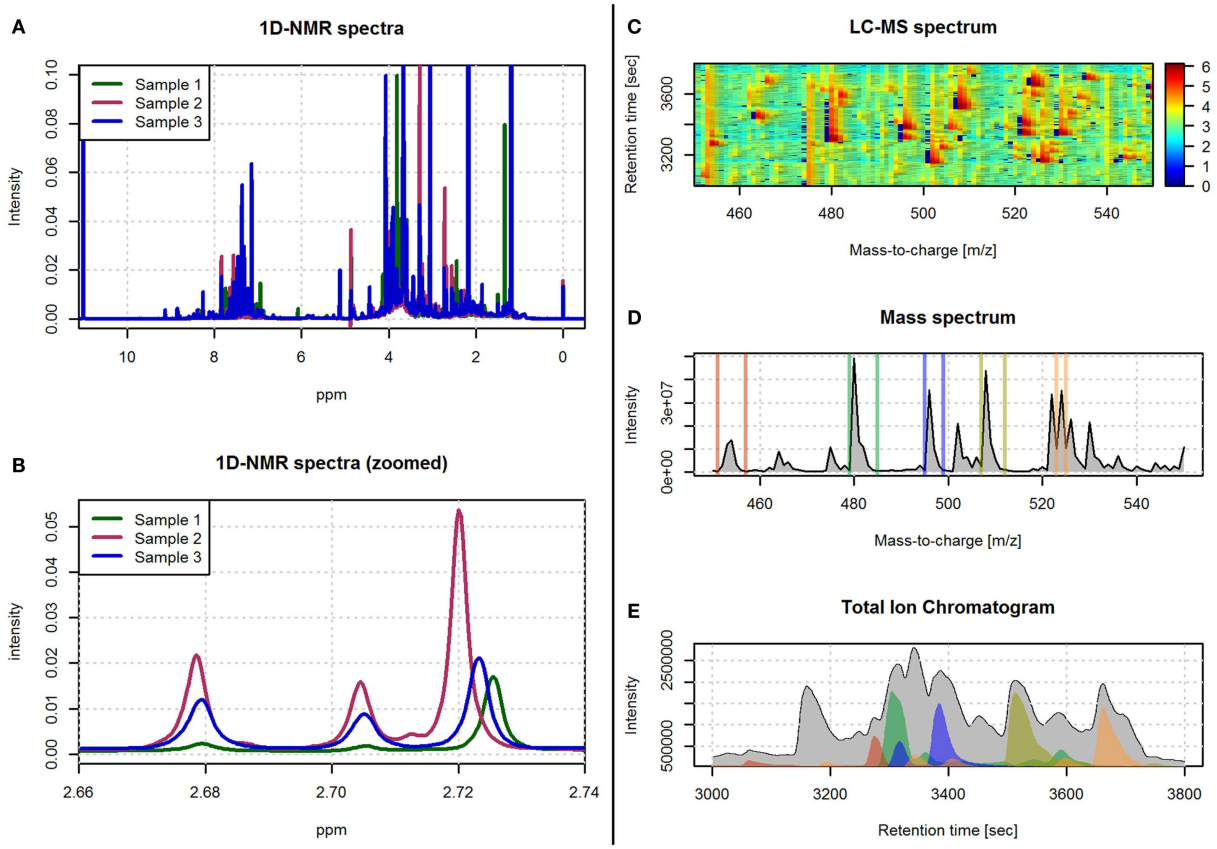


FIGURE 1 | Examples of spectra obtained with ^{1}H-NMR and LC-MS technologies. **(A)** An example of three spectra obtained with 1D ^{1}H-NMR. **(B)** A zoomed view of the spectra in **(A)** in the 2.66–2.74 ppm range. **(C)** An example of a LC-MS spectrum with color-coded intensity and referred by the m/z and retention time axes. **(D)** The sum of the LC-MS spectrum across the m/z axis. **(E)** The total ion chromatogram (i.e., sum of the LC-MS spectrum across the retention time axis). The colored regions in **(E)** correspond to the sum of the LC-MS spectrum limited to the m/z ranges depicted with the same color in **(D)**.

with 1D-NMR spectra. The second dimension in 2D-NMR allows to separate otherwise overlapping spectral peaks and, therefore, gives additional and important information on the chemical properties of the metabolite (Ward et al., 2007). Although 2D-NMR generates a large number of different spectra, these can be globally classified into homonuclear (i.e., ^{1}H–^{1}H-NMR) and heteronuclear (i.e., ^{1}H–^{13}C or ^{1}H–^{15}N) spectra (Marion, 2013). There are also different pulse sequences used to generate the 2D-NMR spectra such as correlation spectrometry (COSY), total correlation spectroscopy (TOCSY), and nuclear Overhauser effect spectroscopy (NOESY). 1D- and 2D-NMR frequency axes are usually referenced by the chemical shift expressed in parts per million (ppm). The chemical shift is calculated as the difference between the resonance frequency and that of a reference substance, subsequently divided by the operating frequency of the spectrometer (Blümich and Callaghan, 1995).

Mass spectrometry is an analytical technique that acquires spectral data in the form of a mass-to-charge ratio (m/z) and a relative intensity of the measured compounds. For the spectrometer to generate the peaks signals for each metabolite, the biological sample first needs to be ionized. The resulting ionized compounds from each molecule will then generate different peak patterns that define the fingerprint of the original molecule. A wide range of instrumental and technical variants are currently available for MS spectrometry. These variants are mainly characterized by different ionization and mass selection methods (El-Aneed et al., 2009). In metabolomics, MS is generally preceded by a separation step. This step reduces the high complexity of the biological sample and allows the MS analysis of different sets of molecules at different times. Liquid and gas chromatography columns (LC and GC, respectively) are the most commonly used separation techniques (Theodoridis et al., 2011). This chromatographic separation technique is based on the interaction of the different metabolites in the sample with the adsorbent materials inside the chromatographic column. This way, metabolites with different chemical properties will require different amounts of time to pass through the column. The time that each metabolite requires, called retention time, is used together with the m/z MS values to generate the two axes of the LC-MS and GC-MS spectral data (**Figure 1**).

In the present review, we will describe the processing and analysis workflows that are commonly used in high-throughput untargeted metabolomic studies. Untargeted metabolomic studies are characterized by the simultaneous measurement of a large number of metabolites from each sample. This strategy, known as top-down strategy, avoids the need for a prior specific hypothesis on a particular set of metabolites and, instead, analyses the global metabolomic profile. Consequently, these studies are characterized by the generation of large amounts of data. This data is not only characterized by its volume but also by its complexity and, therefore, there is a need for high performance bioinformatic tools. Conversely, targeted metabolomic studies are hypothesis-driven experiments and are characterized by the measurement of predefined sets of metabolites with a high level of precision and accuracy. This low level of metabolite analysis is not in the scope of this review, and interested readers are referred to other excellent specific reviews (Roberts et al., 2012; Putri et al., 2013).

In **Figure 2**, we show the typical methodological pipeline of an untargeted metabolomic study. This methodological pipeline starts with the processing of the spectral data to generate the sample metabolic information (i.e., metabolic features). The different methods available to process spectral data are revised in Section "Spectral Processing." Together with metabolite-identification methods, spectral processing methods are highly dependent on the analytical technique used (e.g., NMR, LC-MS, or GC-MS). Once the complete set of metabolomic features has been generated, univariant and multivariant data analysis methods can be applied to investigate: (a) the general structure of the metabolomics data in the dataset and (b) how the different metabolic features are related with the phenotypic data associated with the samples. These analysis methods are reviewed in Section "Data Analysis." The analysis of metabolomic data can often be used to build models that attempt to describe the observed data. Section "Biomarker Discovery in Metabolomics" of the present review describes the different strategies for assessing the performance of these models. In Section "Metabolite Identification and Spectral Databases," we address the important technical issue that is the identification of the metabolites underlying the observed metabolic features (i.e., peak areas and spectral bins). The bioinformatic methods that are actually available for the integration of metabolomic data according with biological knowledge are reviewed in Section "Pathway and Network Analysis of Metabolomic Data." Finally, the different methodologies that allow the integration of metabolomics data with other omics data (e.g., genomics or transcriptomics) are reviewed in Section "Integration of Omics Data." **Table 1** shows a list of the freely available tools that are most commonly used in metabolomic analysis. These tools provide different methodological options for spectral processing, data analysis, or pathway analysis.

SPECTRAL PROCESSING

Spectral processing is a methodological approach aimed at accurately identifying and quantifying the features in the sample spectra of a metabolomics study (**Figure 3**). Metabolomic spectra are sequentially or jointly processed until a final set of feature quantifications is obtained. Spectral processing is also necessary to guarantee that each final measurement will refer to the same metabolomic feature in all samples. The data resulting from spectral processing is generally arranged in a feature quantification matrix (FQM) that contains the quantification of the metabolic features of all the analyzed samples and that will be used as input for subsequent statistical analysis.

SPECTRAL PRE-PROCESSING

In order to improve the signal quality and reduce possible biases present in the raw data, several pre-processing steps are usually applied. In NMR- and MS-based spectra, baseline correction is used to remove low frequency artifacts and differences between

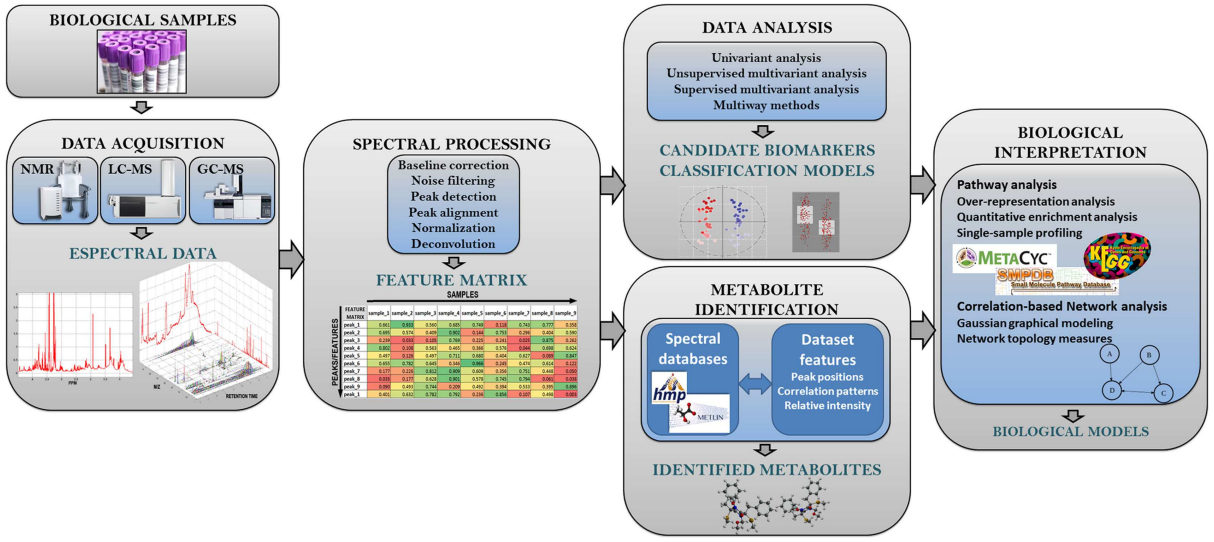


FIGURE 2 | Analysis workflow in untargeted metabolomic studies. This figure shows the different steps of the metabolomic analysis pipeline.

Table 1 | List of tools available for metabolomics spectral processing and data analysis.

Tool	Type	Target	Features[a]	Website	Reference
MetaboAnalyst2	Web	MS and NMR	1–7	http://www.metaboanalyst.ca/	Xia et al. (2012)
XCMS	R	MS	1–3	http://metlin.scripps.edu/xcms/	Smith et al. (2006)
MetSign	MatLab	MS	1–3	http://metaopen.sourceforge.net/	Lommen and Kools (2012)
XCMS online	Web	LC-MS	1–4	https://xcmsonline.scripps.edu/	Tautenhahn et al. (2012b)
MAVEN	Application	LC-MS	1–7	http://genomics-pubs.princeton.edu/mzroll	Melamud et al. (2010)
mzMine2	Application	LC-MS	1–5	http://mzmine.sourceforge.net/	Pluskal et al. (2010)
MAIT	R	LC-MS	1–5	http://b2slab.upc.edu/software-and-downloads	Fernández-Albert et al. (2014)
OpenMS	Application	LC-MS	1–3	http://open-ms.sourceforge.net/	Sturm et al. (2008)
Metabolome express	Web	GC-MS	1–5	https://www.metabolome-express.org/	Carroll et al. (2010)
Metabolite detector	Application	GC-MS	1–4	http://md.tu-bs.de/	Hiller et al. (2009)
MetDAT	Web	MS	1–5	http://smbl.nus.edu.sg/METDAT2/	Biswas et al. (2010)
FOCUS	MatLab	NMR	1–4	http://www.urr.cat/FOCUS/	Alonso et al. (2013)
Automics	Application	NMR	1–2, 5	https://code.google.com/p/automics/	Wang et al. (2009)
Bayesil	Web	NMR	1–4	http://bayesil.ca/	Ravanbakhsh et al. (2014)
Speaq	Application	NMR	1–2, 5	https://code.google.com/p/speaq/	Vu et al. (2011)
MetaboLab	Application	NMR	1–2, 5	http://www.nmrlab.org.uk/	Ludwig and Gunther (2011)
rNMR	R	NMR	8	http://rnmr.nmrfam.wisc.edu/	Lewis et al. (2009)
MetaboMiner	Application	NMR	8	http://wishart.biology.ualberta.ca/metabominer/	Xia et al. (2008)
Muma	R	–	5	http://cran.r-project.org/web/packages/muma	Gaude et al. (2013)
MetaXCMS	R	MS and NMR	5	http://metlin.scripps.edu/metaxcms/	Tautenhahn et al. (2010)
BATMAN	R	NMR	3–4	http://batman.r-forge.r-project.org/	Hao et al. (2012)
AStream	R	LC-MS	4	http://www.urr.cat/AStream/AStream.html	Alonso et al. (2011)
Camera	R	LC-MS	4	http://metlin.scripps.edu/xcms/	Kuhl et al. (2011)
MetaboHunter	Web	NMR	4	http://www.nrcbioinformatics.ca/metabohunter/	Tulpan et al. (2011)
MetScape	Application	–	6–7	http://metscape.ncibi.org/	Gao et al. (2010)
IMPaLA	Web	–	6–7	http://impala.molgen.mpg.de/	Kamburov et al. (2011)
MetExplore	Web	–	6–7	http://metexplore.toulouse.inra.fr/	Cottret et al. (2010)
MetPA	Web	–	6–7	http://metpa.metabolomics.ca/	Xia and Wishart (2010a)
Cytoscape	Application	–	7	http://www.cytoscape.org/	Smoot et al. (2011)
Vanted	Application	–	7	http://vanted.ipk-gatersleben.de/	Rohn et al. (2012)
Paintomics	Web	–	7	http://www.paintomics.org/	García-Alcalde et al. (2011)

This table provides a complete and updated list of the open-source software that is commonly used in the untargeted analysis of metabolomic data.
[a]This column refers to the features included in the tool: spectral pre-processing (1), spectral/peak alignment (2), peak detection (3), metabolite identification (4), data analysis (5), pathway analysis (6), pathway visualization (7), and 2D-NMR analysis (8).

samples that are generated by experimental and instrumental variation (Dietrich et al., 1991; Smith et al., 2006; Xi and Rocke, 2008; Zhang et al., 2010). After this, the application of high-frequency filters may be necessary to remove the electronic noise present in the data that is generated by the measurement equipment.

FEATURE-DETECTION

The objective of the feature-detection step is to identify and quantify the features present in the spectra. Peak-based methods are the most common algorithmic choice for feature-detection in MS-based studies (Gika et al., 2014; Niu et al., 2014; Rafiei and Sleno, 2015). These methods detect the peaks across the spectrum and integrate their areas to provide a quantification of the underlying metabolite. In this approach, spectral alignment is also generally applied either before or after peak detection. In NMR studies, binning-based approaches have been commonly used to detect feature peaks in complex biological samples. However, these methods perform poorly compared to peak-based methods, particularly in those cases where there is significant spectral unalignment, or in those cases where multiple peaks from different metabolites are captured by the same spectral bin (Vu and Laukens, 2013). For these reasons, peak-based methods are increasingly being used in NMR-based studies (Wishart, 2008). Nonetheless, there have been recent developments in binning algorithms, particularly in the detection of the optimal binning boundaries that have improved the performance of this feature-detection approach (Sousa et al., 2013).

Peak overlap is also a common problem in NMR-based studies. Overlapping peaks are treated as one same feature both in binning and peak-based approaches. Consequently, the results obtained from the analysis of these variables can be often hard to interpret. To attempt to solve this problem, spectral deconvolution methods have been developed (Hao et al., 2014). These methods, which are based on the fitting to metabolite spectral templates, are able to extract independent metabolite quantifications from a set of overlapping peaks. The main disadvantage of this

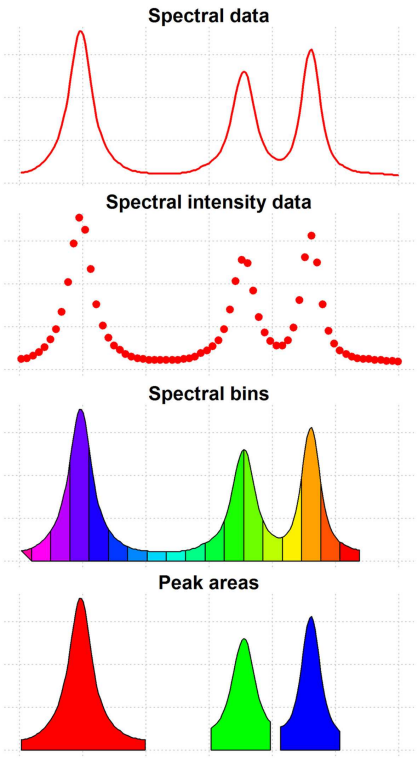


FIGURE 3 | Features of spectral data. This figure shows the different types of features that can be extracted from spectral data and used for data analysis.

type of algorithms, however, is that they depend on the existence of spectral libraries of each metabolite and, therefore, they are unable to quantify peaks arising from previously uncharacterized metabolites.

Peak detection

The most commonly used peak detection algorithms analyze each sample spectrum independently (Tautenhahn et al., 2008, 2012b; Pluskal et al., 2010). These methods are based on two analytical steps (Yang et al., 2009). In the first step, the spectra are smoothed. For this objective, multiple different filters are available (i.e., moving average, Gaussian, Savitzky-Golay. . .; Yang et al., 2009). From these, however, the Wavelet transform-based filters have demonstrated a superior performance, although at the expense of a higher computation time (Du et al., 2006; Tautenhahn et al., 2008). This performance improvement is mainly due to the ability of the Wavelet transform to work with the unequal peak widths that characterize metabolomic spectra. In the second step, the different metabolite peaks are identified using one or multiple detection thresholds. These thresholds are applied to different parameters such as the signal-to-noise ratio, the intensity, or the area of each peak from the resulting filtered spectra (Yang et al., 2009). In metabolomics studies involving large numbers of samples, a frequency filter (i.e., consensus peak signal), can be also applied so that only those peaks that are present in a minimum percentage of samples are selected for downstream analysis.

Spectral alignment

Spectral alignment is one of the main processing steps in metabolomic studies involving multiple samples. When analyzing multiple spectra, the position of the peaks corresponding to the same metabolic feature may be affected by non-linear shifts. In NMR-based studies, these shifts are observed in the ppm axis and are usually introduced by differences in the chemical environment of the sample like ionic strength, pH, or protein content (Weljie et al., 2006; Xiao et al., 2009). In MS-based studies, peak shifts are mainly observed across the retention time axis, and are generally associated with changes in the stationary phase of the chromatographic column (Burton et al., 2008). Spectral alignment methods must be therefore applied to correct this undesired variability in the samples that can profoundly affect the quality of the study. Spectral alignment algorithms can be divided in two main groups: (i) spectral alignment methods, where the spectral data is aligned before peak detection and (ii) peak-based alignment methods, where spectral peaks are aligned across samples once they have been detected using their coordinates (ppm in NMR, and m/z and retention time in LC/GC-MS).

Spectral alignment methods are classified into warping and segmenting methods. Warping methods are based on the application of a non-linear transformation to the ppm (in NMR spectra) or the retention time (in LC/GC-MS) axis in order to maximize the correlation between the spectra. The alignment is then performed by either stretching or shrinking spectral segments to reach this correlation maximization. Among these methods, correlation optimized warping (COW) and dynamic time warping (DTW) are the most commonly used. COW is a segmental alignment method that aligns one sample spectrum toward a reference spectrum (Tomasi et al., 2004). This is done by splitting the original sample and reference spectra into small segments, and by separately aligning each pair of segments. Alignment is performed through dynamic programing in such a way that limited changes in segment lengths are allowed. This way, the overall correlation between both spectra is effectively maximized. In the particular case of crowded spectral regions with large peak shifts, COW has demonstrated to perform particularly well compared to other methods. An alternative to COW method, DTW is a spectral alignment method (Tomasi et al., 2004) that is also based on dynamic programing, and where a warping path is computed to which the connected data points of each spectrum are equivalent. During this last decade, other warping approaches have been developed (Eilers, 2003; Forshed et al., 2003; Lee and Woodruff, 2004; Clifford et al., 2009).

Spectral segmenting methods differ from spectral warping methods in that alignment is performed by applying a constant shift to all the spectral points. These methods either align the overall spectra or split the spectra into smaller segments and independently align each resulting segment. The Icoshift algorithm (Savorani et al., 2010) is one of the most commonly used segmentation methods, and is based on the convergence toward a reference signal. This convergence is performed by applying shifts that maximize the segment spectral correlation, which is normally computed using the fast Fourier transform (FFT) to

speed up the required calculations (Wong et al., 2005). Icoshift and other correlation-based methods can also be combined with automatic segmentation methods (Veselkov et al., 2008), which are able to optimally split the spectra in order to improve the alignment of the resulting spectral segments. However, the use of a reference spectrum has several disadvantages. Very recently, the RUNAS algorithm implemented in the FOCUS processing workflow (Alonso et al., 2013) has provided a spectral segmenting method that avoids the use of a reference spectrum. Instead, the FOCUS method uses the information from the different sample spectra to iteratively maximize the inter-sample weighted-mean correlation. This approach has shown that avoiding the use of a reference spectrum is a powerful strategy to avoid many of the analytical biases derived from its use. These biases are mainly due to the fact that the reference spectrum may not be representative of the spectral diversity present in the samples. FOCUS alignment algorithm has also shown that an appropriate spectral transformation prior to alignment avoids the biases due to the presence of multiple peaks in the same alignment window. Under these conditions, the methods based in correlation maximization without prior transformation are more prone to align the most relevant peak of each sample regardless of whether they correspond to the same metabolic feature or not.

Fast Fourier transform-based segmenting methods such as RAFFT, Icoshift, and FOCUS not only are able to process large metabolomics datasets in a reduced amount of time, but also have shown to perform better than spectral warping methods (Giskeødegård et al., 2010; Savorani et al., 2010; Alonso et al., 2013; Jiang et al., 2013). Within the different segmenting methods, reference-free methods avoid the biases introduced by using reference spectra, but at a cost of being more computationally intensive.

Of relevance, the results reported by several performance comparison studies using either NMR or MS have demonstrated that spectral alignment algorithms have a good performance irrespective of the analytical technique that has been used (MS or NMR; Van Nederkassel et al., 2006; Giskeødegård et al., 2010). Consequently, methods that were initially developed to align NMR spectra are also applied to align MS spectra and vice versa.

Compared to the warping and segmentation alignment methods, peak-based methods are applied after peak detection. In these methods, peak coordinates are used to perform the alignment. This type of method is implemented in the XCMS software (Tautenhahn et al., 2012b), one of the most commonly used methods to process data from LC-MS studies. Given that the shifts along the m/z axis are minimal and the m/z axis has a high resolution, the data can be safely binned in m/z intervals, and peak alignment performed on each bin along the chromatographic time. The XCMS algorithm computes the retention time boundaries within which the observed peaks are expected to represent the same metabolomic feature across the different samples. The computation of these retention time boundaries is performed by using a kernel density estimator. Another common alignment method used in MS is the RANSAC algorithm (Pluskal et al., 2010). In this approach, the corresponding peaks across samples are identified by using a LOESS regression on different retention times and m/z windows.

FEATURE NORMALIZATION

In order to perform an accurate quantification of the features in a metabolomic analysis, a data normalization step is generally required. The objective of normalization is to remove undesired systematic biases, so that only biologically relevant differences are present in the data. This procedure is crucial when analyzing complex biofluids like blood, where the differences in metabolite concentration between samples can be high, and the introduction of internal standards is complicated. Although multiple statistical models have been developed for this objective (Craig et al., 2006; Kohl et al., 2012), the two perhaps most commonly used methods are the use of endogenous stable metabolites (like creatinine in urine) and the use of the total spectral area [i.e., area under the curve (AUC); Weljie et al., 2006; Rasmussen et al., 2011].

DECONVOLUTION METHODS IN TARGETED ANALYSIS

One of the main limitations for the quantification of metabolomic features is the overlap between peaks from different metabolites. NMR and GC-MS spectra are particularly prone to this type of bias. In order to deal with this technical issue, several methodological approaches have been developed. These approaches are based on spectral deconvolution (Chylla et al., 2011; Astle et al., 2012; Du and Zeisel, 2013; Hao et al., 2014), a signal processing technique that estimates the relative area corresponding to each individual peak when multiple peaks overlap within the same spectral region (**Figure 4**). However, an important limitation of deconvolution methods is that prior knowledge of the compounds

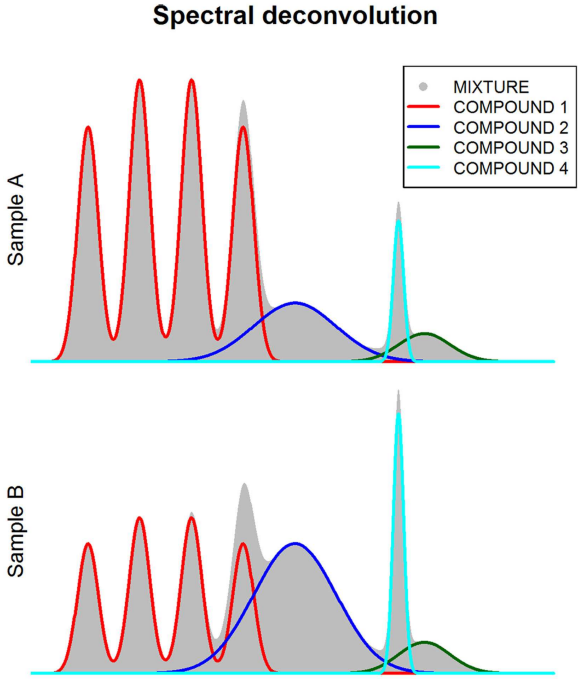


FIGURE 4 | Spectral deconvolution. This figure shows how spectra (gray shaded area) can be decomposed (i.e., deconvoluted) in multiple components corresponding to different metabolite compounds.

present in the mixture is required. Additionally, the use of these methods in untargeted metabolite studies is yet not possible due to computational intractability (Hao et al., 2014).

The usual input data for these methods is the spectral data from the study and at template library containing the reference peak patterns of each metabolite. Currently, there are multiple methods available for spectral deconvolution of NMR data (Chylla et al., 2011; Zheng et al., 2011; Astle et al., 2012; Hao et al., 2014) and they are mostly based on Bayesian model selection. Among them, BATMAN (Hao et al., 2012) is one of the most frequently used, providing a rich and user-friendly interface and a complete protocol to perform this type of analysis (Hao et al., 2014). BATMAN is an open-source software and its performance has been demonstrated to be very similar to that of the NMR Suite software package (Chenomx Inc., Edmonton, AB, Canada; Weljie et al., 2006), a proprietary software that is considered a gold standard for NMR metabolomics (Chenomx Inc., Edmonton, AB, Canada; Weljie et al., 2006). The NMR Suite itself provides a semi-automated tool for spectral deconvolution which allows interactive fitting of the metabolite peaks to the reference metabolite spectra. The major disadvantage of this tool is the large amount of time required to process large sample datasets and the need of highly skilled data analysis specialists.

GC-MS methods for spectral deconvolution are mostly based on unsupervised approaches that do not require the prior knowledge of the compounds presents in the sample (Stein, 1999; Hiller et al., 2009; Ni et al., 2012). These approaches are mainly based on three steps, namely: (a) noise analysis for selecting the spectral segments to be deconvoluted, (b) component perception for identification of the individual components present in each segment, and (c) deconvolution for fitting the individual components to the overall spectral shape. Du et al. provide an extensive review of these methods (Du and Zeisel, 2013).

DATA ANALYSIS

Once the metabolite features are robustly quantified, there are multiple univariate and multivariate statistical methods that can be used to perform the desired study analysis. These groups of techniques are commonly known as chemometric methods (Madsen et al., 2010) and usually require some degree of expertise to be conveniently applied. In the following sections, we define the most commonly used metabolomic features, and we describe the most commonly used chemometric methods.

METABOLOMIC FEATURES

After applying the adequate pre-processing methods, metabolomics data is usually reduced to a FQM. In this data representation, rows correspond to the samples and columns correspond to the different metabolomic features. Each metabolomic feature is intrinsically related to the concentration of a particular metabolite. Depending on the analytical technique and the spectral processing workflow that have been used, different metabolomic features are used as input for data analysis (**Figure 3**):

- Spectral peak areas: one of the most commonly used features in high-throughput metabolomics data (NMR-based or MS-based). They are computed through the integration of the peaks identified and aligned using the methods described in the previous section (see Spectral Pre-Processing to Deconvolution Methods in Targeted Analysis). Once this data has been analyzed, the identification of the metabolites representing the relevant peaks is required in order to provide biological meaning to the results. Metabolite-identification methods are reviewed in Section "Metabolite Identification and Spectral Databases."
- Metabolite concentrations: in contrast to the previous features, metabolite identification can be performed prior to data analysis in order to obtain absolute or relative metabolite concentrations to be used as input for data analysis (Wishart, 2008; Zhou et al., 2012). This type of features allow both to reduce the high redundancy of peak areas (i.e., one metabolite is often represented by multiple spectral peaks), and to provide biological significance to all the analyzed features. The most common metabolite-identification methods are reviewed in Section "Metabolite Identification and Spectral Databases."
- Spectral bin areas: in addition to peak areas and metabolite concentrations, spectral bins (or also buckets) are also commonly used features in NMR-based studies. This technique consists of dividing the spectra into evenly spaced regions that are later integrated to obtain the corresponding spectral bin areas. In order to mitigate problems such as peaks lying in two consecutive integration regions, some methods have implemented uneven binning algorithms like dynamic adaptative binning (Anderson et al., 2011), Gaussian binning (Anderson et al., 2008), and adaptative intelligent binning (De Meyer et al., 2008). This feature estimation approach has, however, some inherent disadvantages produced by the presence of uninformative features in the spectra (i.e., spectral areas without spectral peaks) and the lack of inter-sample feature correspondence when spectra are heavily affected by unalignment (e.g., urine samples with large pH variability).

UNIVARIATE ANALYSIS METHODS

Univariate methods analyze metabolomic features independently. They are common statistical analysis approaches and, therefore, their main advantage is their ease of use and interpretation. However, their main disadvantage is that they do not take into account the presence of interactions between the different metabolic features. The metabolomic data obtained from biological samples is often very complex with the presence of correlations between features from the same metabolite and correlations between metabolites from the same pathway. Also, the effect of potential confounding variables like gender, diet, or body mass index is not taken into account by these analysis methods, increasing the probability of obtaining false positive or false negative results (Winnike et al., 2009; Rasmussen et al., 2011; Townsend et al., 2013).

Several univariate analysis methods are available for metabolomic data analysis. The selection of the method will depend on the statistical properties of the feature distribution (Broadhurst and Kell, 2006; Vinaixa et al., 2012). For example, when assessing differences between two or more groups, parametric tests such as Student's *t*-test and ANOVA are commonly applied, provided that normality assumptions are conveniently verified. The latter can be confirmed using the Kolmogorov–Smirnov normality test or Bartlett's homogeneity of variances

test. In those cases where normality of the data cannot be assumed, non-parametric tests such as Mann–Whitney *U* test or Kruskal–Wallis one-way analysis of variance are preferable.

In addition to choose the most appropriate statistical analysis test, another important consideration in metabolomic data analysis is the multiple testing problem. In most metabolomic studies, a large number of metabolomic features are analyzed simultaneously and, therefore, the probability of finding a statistically significant result by chance (i.e., false positive) is high. In order to control for this multiple testing issue, several correction methods are available. Each method is characterized by a particular balance between avoiding false metabolite associations (i.e., false positives) and prevents discarding true associations (i.e., false negatives). Depending on the study design, researchers might decide to use a more or less conservative approach. The Bonferroni correction is perhaps the most conservative multiple test correction approach, where the number of type I errors (false positives) regarding to the total number of hypotheses tested [i.e., defined as familywise error rate (FWER)] is minimized at the expense of increasing type II errors (false negatives). In the Bonferroni correction, the significance level for one hypothesis (i.e., alpha value), is divided by the number of hypotheses tested simultaneously. Although a very conservative approach, especially when the hypotheses tested are not independent, many researchers advocate its use in metabolomic studies (Broadhurst and Kell, 2006). Recently, Chadeau-Hyam et al. assessed the metabolome-wide significance level (MWSL) for biomarker identification in urine using a permutation-based method to estimate the correct FWER (Chadeau-Hyam et al., 2010). Their method took into account metabolite collinearity and reported that a conservative estimate of the independent number of tests is 35% of the performed tests. This result indicates that the Bonferroni multiple test correction method might be over conservative.

Other less conservative multiple test correction methods are however available and are mostly based on the minimization of the false-discovery rate (FDR; Benjamini and Hochberg, 1995). While Bonferroni and other FWER-based methods minimize the probability of at least one false positive in the overall set of tests, FDR-based methods minimize the expected proportion of false positives on the total number of positives (Van Den Oord, 2008). Most of these methods have been extensively used for the analysis of gene-expression microarray data, where thousands of genes are tested in parallel (Reiner et al., 2003; Jung, 2005; Xie et al., 2005). In untargeted metabolomic studies, where large numbers of metabolites are simultaneously analyzed, and where it is also expected that more than one or two of these biomarkers will be associated, the use of less strict multiple correction methods like FDR methods might be more useful.

MULTIVARIATE ANALYSIS METHODS

In contrast to univariate methods, multivariate analysis methods take into account all the metabolomic features simultaneously and, consequently, they can identify relationship patterns between them. These pattern-recognition methods can be classified into two groups: supervised and unsupervised methods. In unsupervised analysis methods, the similarity patterns within the data are identified without taking into account the type or class of the study samples. In supervised methods, the sample labels are used in order to identify those features or features combinations that are more associated with a phenotype of interest. Supervised methods are also the basis for building prediction models.

Unsupervised methods

Unsupervised methods are often applied to summarize the complex metabolomic data. They provide an effective way to detect data patterns that are correlated with experimental and/or biological variables. Principal component analysis (PCA) is the most commonly used unsupervised method in metabolomic studies (Wold et al., 1987; Bro and Smilde, 2014). PCA is based on the linear transformation of the metabolic features into a set of linearly uncorrelated (i.e., orthogonal) variables known as principal components. This decomposition method maximizes the variance explained by the first component while the subsequent components explain increasingly reduced amounts of variance. At the same time, PCA minimizes the covariance between these components (i.e., they are independent of each other). After applying the PCA method, a set of loading vectors and score vectors are obtained. The loading vectors represent the principal components, and each vector coefficient corresponds to the individual contribution of each variable to the principal component. The score vectors represent the projection of each sample onto the new orthogonal basis. Plotting these sample scores over the first principal components is a convenient way of summarizing the global dataset, since normally these first principal components capture most of the variability in the dataset. PCA is also used in metabolomics studies to assess data quality, since it can identify sample outliers or reveal hidden biases in the study. For example, PCA has been used in several studies to determine the impact of technical variation in the analysis of metabolic profiles (Gika et al., 2008; Winnike et al., 2009; Rasmussen et al., 2011; Yin et al., 2013).

Other unsupervised methods like hierarchical clustering analysis (HCA) and self-organizing maps (SOMs) have also been applied to metabolomic data. These methods can be particularly suitable to detect non-linear trends in the data that are not conveniently covered by PCA. SOMs have been used in metabolomics studies to visualize metabolic phenotypes and feature patterns as well as to prioritize the metabolites of interest based on their similarity (Kohonen et al., 2000; Meinicke et al., 2008; Mäkinen et al., 2008; Goodwin et al., 2014). HCA is also a powerful clustering and visualization tool that provides a clustering procedure at the feature and sample levels according to a predefined distance measure (Brauer et al., 2006; Sreekumar et al., 2009).

Supervised methods

Supervised methods are used to identify metabolic patterns that are correlated with the phenotypic variable of interest while downweighting the other sources of variance. These methods are also the basis for building classifiers based on metabolomic features (Xia et al., 2013). Partial least squares (PLS; Fonville et al., 2010) is one of the most widely used supervised method in metabolomics. It can be used either as a regression analysis (i.e., quantitative variable of interest) or as a binary classifier (PLS-DA; i.e., binary variable of interest). Unlike PCA, PLS components do not maximize the explained dataset variance but the covariance between

the variable of interest and the metabolomic data. Therefore, the feature coefficients (loadings) of PLS components represent a measure of how much a feature contributes to the discrimination of the different sample groups. However, one weakness of PLS is that some metabolic features that are not correlated with the variable of interest can influence the results. In order to deal with this problem, orthogonal PLS (O-PLS; Trygg and Wold, 2002) were developed. O-PLS models evolved from PLS models and factorize the data variance into two components: a first component which is correlated with the variable of interest and a second uncorrelated component (i.e., orthogonal). Classification of metabolomics samples is commonly performed by fitting the discriminant analysis versions of PLS and O-PLS models (i.e., PLS-DA, O-PLS-DA; Kemsley, 1996; Bylesjö et al., 2006).

The performance of PLS and O-PLS models has been extensively compared but, to date, there is no agreement as to which of the two methods is superior (Tapp and Kemsley, 2009). In the last years, however, a progressive move from the use of PLS models to O-PLS models has been observed in the metabolomics field (Fonville et al., 2010).

Support vector machines (SVMs) are another class of supervised analysis methods to build classifiers based on metabolomic data (Mahadevan et al., 2008; Kim et al., 2010; Luts et al., 2010). Although classifiers based on SVM are harder to interpret, they are able to manage the presence of non-linear relations between the metabolomic data and the variable of interest.

Multiway methods for longitudinal metabolomic data

There is also a wide range of methods that are designed to provide a comprehensive interpretation of the metabolic changes according to the organization of the analyzed samples (i.e., samples from different tissues or corresponding to time series in a longitudinal study). These methods decompose the original multiway (i.e., multi-dimensional) data matrix into a set of easily interpretable factors. In NMR studies, two of the most commonly used methods are parallel factor analysis (PARAFAC) and multivariate curve resolution (MCR). The input data for these methods is commonly a three dimensional (3D) matrix with coefficients c_{ijk} (where i represents a metabolic feature, j the analyzed individual, and k the tissue from which the sample was extracted or the sample extraction time-point). The PARAFAC analysis of a 3D matrix generates three loading matrices that capture the contributions of each metabolic feature, of each individual, and of each tissue type or time-point. Alternatively, MCR analysis decomposes the 3D matrix into a set of two factors which contain the contributions of each metabolic feature and each analyzed sample. To do this, the 3D matrix must be fitted in a 2D matrix, where the different metabolic features are arranged on the first dimension while the each individual and tissue/time-point are arranged on the second dimension (Peré-Trepat et al., 2007; Karakach et al., 2009; Montoliu et al., 2009; Martin et al., 2010).

BIOMARKER DISCOVERY IN METABOLOMICS

One of the most promising applications of metabolomics in the medical sciences is the identification of biomarkers. New metabolomic biomarkers are usually determined using supervised analysis models since they are capable to aggregate the evidence of multiple metabolites. The usefulness of the resulting classification models must be then evaluated in order to consider their use in real clinical settings. Performance assessment and model validation are crucial analytical steps for the evaluation of metabolomic classification models.

PERFORMANCE ASSESSMENT

Performance assessment measures how well the outcome predicted by our model matches the real outcome. Several complementary measures are available to assess the classifier performance: predictive accuracy (percentage of correctly classified subjects), sensitivity (percentage of true positives that are correctly classified), and specificity (percentage of true negatives that are correctly classified). These three measures allow the assessment of the classifier performance given a fixed decision boundary. However, these performance measures tend to be dependent on the outcome prevalence and on the decision boundary chosen (Xia et al., 2013). The receiver operating characteristic (ROC) curve avoids this type of bias and is the most used performance assessment method. ROC curve estimation is a non-parametric procedure consisting of the comparison of specificity against sensitivity according to a specific decision boundary. ROC curves are often summarized by the AUC metric. The AUC metric gives the probability that a classifier will rank a randomly chosen positive sample higher than a randomly chosen negative one. Therefore, a perfect classifier will obtain AUC = 1 while a random classifier will obtain AUC close to 0.5. An AUC >0.7 is often considered the minimal performance for a biomarker test to be considered clinically useful (Xia et al., 2013). In addition to the overall performance assessment using the AUC metric, the ROC curves can also be used to determine the optimal decision boundary for the classifier (Xia et al., 2013). ROC curve estimation is a common analysis and therefore, multiple tools are available for ROC-based performance evaluation like the R packages ROCR (Sing et al., 2005) and pROC (Robin et al., 2011), as well as the ROCCET (Xia et al., 2013) web application.

MODEL VALIDATION

When designing classification models, a validation step is required to estimate how well the classification model will perform when applied to new samples. This step is particularly important when using small sample sizes in order to discard model overfitting. Two main approaches are available for performing this task: permutation testing and cross-validation (Westerhuis et al., 2008).

The aim of the permutation-based validation is to measure the performance of the predictor model by determining the probability of observing an equal or better performance by pure chance. This analysis is performed by estimating the null distribution of the performance measures (i.e., AUC) under the assumption that no differences exist between sample groups. This is done by randomly permuting multiple times the sample group classes (e.g., case-control) and calculating the statistic under each permuted dataset. Once computed, the performance measures of the *true* model (i.e., based on the real sample status) should lie outside the chosen confidence intervals (e.g., 95 or 99%) of the estimated null distributions in order to be considered significant. In contrast with the permutation approach, cross-validation approaches estimate the predictive performance of a classifier using an iterative

approach. At each round of cross-validation, the total sample is split into a training group and a testing group. In the former group, the predictor model is built using a specific set of parameters. The performance of this model is then evaluated using the remaining group of samples. This procedure is repeated several times so that all the samples have been used once as a testing group. Averaging these results we will obtain an unbiased estimate of the performance of the predictor. The size of the testing sample can be composed by several samples (i.e., *n*-fold cross-validation) or can be as small as a single individual (i.e., leave-one out cross-validation). This approach provides a good measure of how data overfitting affects to the computed model. When the used models require optimization (i.e., optimal number of PLS/O-PLS components to be used) a double cross-validation schema is usually required: a first cross-validation step is applied to optimize the model and a second step for assessing the model quality (Westerhuis et al., 2008; Szymanska et al., 2012). The double cross-validation schema requires the dataset to be iteratively split in two sets S1 and S2. In the first step, the S1 set is randomly divided into two subsets S11 and S12, where S11 is used to compute models with different number of components and the S12 set is used to evaluate the prediction power of each model. This procedure is repeated until all the samples in S1 have been once in the S12 set, and the model with the lowest prediction error is selected. In the second step, the S2 set is used to assess the performance of the optimal model as computed in step one. This global analysis is performed recursively by randomly splitting the global dataset in sets S1 and S2 until all the samples have been once in S1. Further details on the different types of cross-validations are described in more detail elsewhere (Westerhuis et al., 2008; Szymanska et al., 2012).

METABOLITE IDENTIFICATION AND SPECTRAL DATABASES

Metabolite identification is one of the major challenges of high-throughput metabolomic analysis. This step is indispensable to confer a biological meaning to the associated features in a metabolomic study. In MS-based studies, the common metabolite-identification approach is based on querying metabolomic databases for the neutral molecular mass values of the identified peaks using a tolerance window. The neutral molecular mass is inferred from the peak m/z value, and depends on the chemical nature of the identified peak (i.e., ionization mode and ionization adduct). Assuming no prior knowledge, each peak m/z value can lead to multiple plausible neutral molecular masses that can represent different ionization adducts (H^+, Na^+, K^+, . . .). This multiplicity often results in a high number of false positive identifications. In order to reduce false positives, several methods have been developed. AStream and Camera are methods designed to identify isotopic and adduct patterns in order to reduce data complexity in MS experiments (Alonso et al., 2011; Kuhl et al., 2011). Using these approaches, the chemical nature of each selected ion peak is estimated, and only one neutral mass is inferred from each identified pattern. Using these methods has the added advantage of improving the ascertainment of true biological compounds.

In NMR-based studies, automatic metabolite identification is commonly performed by matching the measured NMR peaks against a set of reference metabolite patterns. Each metabolite reference spectrum is defined by one or multiple peaks, which are characterized by their ppm positions and their relative intensities. MetaboHunter is an online tool for identifying compounds by matching the reference peak positions against the list of detected peak positions (Tulpan et al., 2011). However, this approach can lead to high false positive rates, since it only uses one peak parameter to match reference peaks. The MetaboHunter approach has been superseded by more recent methods based on the valid cluster concept (Mercier et al., 2011; Jacob et al., 2013). In addition to using the ppm position, these methods include peak intensities and inter-sample intensity correlation as parameters for matching data peaks to reference peaks. The NMR analysis workflow implemented in FOCUS follows this same metabolite-identification approach, with the added advantage that it also accounts for the presence of missing peaks generated by spectral overlapping (Alonso et al., 2013).

Metabolite spectral databases are essential for metabolite identification. The quality of the stored data as well as the number of metabolite spectra available in these databases is critical for the performance of identification algorithms. During the last years, multiple databases have been developed (**Table 2**) and the number of available metabolite reference spectra is continuously growing (Ellinger et al., 2013; Fukushima and Kusano, 2013). The Human Metabolome Database (HMDB) is perhaps the most extensive public metabolomic spectral database to date (Wishart et al., 2013). The HMDB stores >40,000 different metabolite entries, with exhaustive biological metadata and MS/NMR spectral references. In addition to spectral databases, several studies have also contributed to characterize the metabolome of multiple types of samples. Many of these reference studies are also exceptional resources of high quality data associated with the biofluid, tissue, or cell type of interest (Wishart et al., 2008; Psychogios et al., 2011; Bouatra et al., 2013).

PATHWAY AND NETWORK ANALYSIS OF METABOLOMIC DATA

Pathway and network analysis approaches increase the information generated by metabolomic studies. Both approaches exploit the relational properties present in metabolomic data. Pathway analysis uses prior biological knowledge to analyze metabolite patterns from an integrative point of view. Alternatively, network analysis uses the high degree of correlation existing in metabolomics data to build metabolic networks that characterize the complex relationships existing in the set of measured metabolites.

PATHWAY ANALYSIS

Until very recently, when analyzing metabolomic data no prior knowledge regarding metabolite relationships could be assumed. During the last years, however, the biological knowledge available for metabolomics studies has been constantly increasing. Metabolic pathways are groups of metabolites that are related to the same biological process, and that are directly or indirectly connected by one or multiple enzymatic reactions. Biological databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al., 2012), small molecule pathway database (SMPDB; Jewison et al., 2014), EHMN (Ma et al., 2007), WikiPathways (Kelder et al., 2012), and MetaCyc (Caspi et al., 2008) provide

Table 2 | Spectral databases available for metabolite identification.

Database	Spectral data	Website	Statistics	Reference
HMDB	MS/NMR	http://www.hmdb.ca	41,806 metabolite entries and 1,579 metabolites with spectra (^{1}H-NMR, LC-MS, GC-MS ...)	Wishart et al. (2013)
LMSD	MS	http://www.lipidmaps.org	37,500 lipid structures with MS/MS spectra	Sud et al. (2007)
METLIN	MS	http://metlin.scripps.edu	240,516 metabolite entries and 12,057 metabolites with MS/MS spectra	Tautenhahn et al. (2012a)
TOCCATA COLMAR	NMR	http://spin.ccic.ohio-state.edu	Multiple spectral NMR datasets: ^{1}H- and ^{13}C-NMR, 2D ^{13}C–^{13}C TOCSY ($n = 463$), 2D ^{1}H–^{1}H TOCSY and ^{13}C–^{1}H HSQC-TOCSY ($n = 475$), and 2D ^{13}C–^{1}H HSQC ($n = 555$)	Robinette et al. (2008), Bingol et al. (2012, 2014, 2015)
MassBank	MS	http://www.massbank.jp	2,337 metabolites and 40,889 spectra (LC-MS, GC-MS ...)	Horai et al. (2010)
Golm metabolome	GC-MS	http://gmd.mpimp-golm.mpg.de	2,019 metabolites with GC-MS spectra	Hummel et al. (2007)
BMRB	NMR	http://www.bmrb.wisc.edu	9,841 biomolecules with ^{1}H, ^{13}C, or ^{15}N spectra	Ulrich et al. (2008)
Madison	NMR	http://mmcd.nmrfam.wisc.edu	794 compounds with spectra including ^{1}H, ^{13}C, ^{1}H–^{1}H, ^{1}H–^{13}C ...	Cui et al. (2008)
NMRShiftDB	NMR	http://nmrshiftdb.nmr.uni-koeln.de	42,840 structures and 50,897 measured spectra	Steinbeck et al. (2003)
RIKEN	MS/NMR	http://prime.psc.riken.jp	1,589 metabolites (*Arabidopsis*)	Akiyama et al. (2008), Sakurai et al. (2013)
Birmingham Metabolite Library	NMR	http://www.bml-nmr.org	208 metabolites and 3,328 1D- and 2D-NMR spectra	Ludwig et al. (2012)

This table shows a list of the spectral databases that are most commonly used in current metabolomics studies to characterize the associated metabolite features.

exhaustive information of a large number of metabolic pathways (**Table 3**). The availability of this data is therefore enabling the use of pathway-based approaches in metabolomics. These methods are currently referred as metabolite set enrichment analysis (MSEA), and are methodologically based on the gene set enrichment analysis (GSEA) approach, designed for pathway analysis of gene-expression data (Khatri et al., 2012).

To date, three different approaches have been developed to perform MSEA (Xia and Wishart, 2010b):

- Overrepresentation analysis (ORA): Given a list of metabolite pathways or groups of metabolites of interest, a hypergeometric test or a Fisher's Exact test is used to evaluate whether the metabolites of these groups are represented more than expected by chance. When the input metabolite list is defined as the set of metabolites which are differentially expressed in the analyzed phenotypes, the ORA results may identify metabolic pathways that are globally associated to these phenotypes.
- Quantitative enrichment analysis (QEA): Unlike ORA, the input data for this method is a set of metabolite concentrations from multiple samples. Enriched pathways can be identified using different approaches like globaltest (Goeman et al., 2004), globalAncova (Hummel et al., 2008), or the Wilcoxon-based test (Adjaye et al., 2005). Enriched pathways include pathways where a few number of compounds are significantly changed or pathways where a large number of metabolites are slightly but consistently changed (Xia and Wishart, 2010b).
- Single-sample profiling (SSP): While the two previous methods are suited for studies involving large numbers of samples, this approach can be used at the sample level. The input data for SSP analysis is an input list of normalized metabolite concentrations in a common biofluid, tissue, or cell type and a database with the normal concentration ranges of these metabolites in the sample. From this input data, SSP identifies the set of metabolites showing levels significantly different from the normal concentration ranges.

In order to improve the interpretability of pathway analysis results, MSEA results can be combined with pathway topological measures. These measures allow the assessment of impact of the unbalanced metabolites within the overrepresented pathway. First, single impacts are evaluated using the degree and betweenness network centrality measures of each metabolite (Aittokallio and Schwikowski, 2006). Subsequently, the overall impact (i.e., pathway impact; Xia and Wishart, 2010a) is calculated as the sum of the single impact measures of the unbalanced metabolites normalized by the sum of the impact measures of all the metabolites within the pathway.

Metabolomics researchers currently have a wide variety of software tools to analyze metabolomic data at the pathway level. Applications such as Paintomics (García-Alcalde et al., 2011), Vanted (Rohn et al., 2012), and Cytoscape (Smoot et al., 2011) provide different pathway visualization tools. In these tools, the metabolites are mapped on predefined metabolic pathways, and

Table 3 | Biological databases for pathway analysis.

Database	Description	Website	Reference
Kyoto Encyclopedia of Genes and Genomes (KEGG)	466 pathways, 17,333 metabolites, and 9,764 biochemical reactions	http://www.genome.jp/kegg/	Kanehisa et al. (2012)
MetaCyc	2260 pathways from 2600 different organisms	http://metacyc.org/	Caspi et al. (2008)
The small molecule pathway database (SMPDB)	1,594 metabolites mapping 727 small molecule pathways found in humans	http://www.smpdb.ca/	Jewison et al. (2014)
WikiPathways	1,910 pathways	http://wikipathways.org/	Kelder et al. (2012)
Plant metabolic network (PMN/PlantCyc)	Multi-species pathway database for plant metabolomics	http://www.plantcyc.org/	Chae et al. (2014)

This table describes the main databases that provide biological information on metabolites and metabolic pathways.

allow a high level of interaction with the data. In addition to visualization tools, Impala (Kamburov et al., 2011) and MetScape2 (Karnovsky et al., 2012) are software tools that also implement specific MSEA methods. Finally, Metaboanalyst is a highly versatile pathway analysis tool, providing a wide range of MSEA methods as well as topological and visualization tools (Xia et al., 2012).

CORRELATION-BASED NETWORK ANALYSIS

One of the main features of biologic data is the high level of correlation existing between the different elements (i.e., mRNAs, proteins, metabolites). Part of these relational patterns is due to metabolites that belong to common metabolic pathways. In other cases, however, the observed correlations may be due to other causes like global perturbations (i.e., metabolic compounds showing diurnal variation in time series analysis), specific perturbations (i.e., changes in enzyme concentrations spread through their related metabolic pathways), or the intrinsic variability of metabolomic data (Steuer et al., 2003; Camacho et al., 2005; Steuer, 2006). Consequently, metabolites that do not show significant differences across the studied phenotypes may still show different correlation patterns with other metabolites in each phenotype. These patterns can provide valuable information about the underlying metabolic network associated to a specific biological process (Steuer, 2006).

Unlike pathway analysis, correlation-based methods build metabolite networks according to the relationship patterns observed in the experiment data. In the resulting network, each metabolite is represented by a network node but, in contrast to pathway analysis, the links between nodes represent the level of mathematical correlation between each pair of metabolites. In metabolomics data, high correlation coefficients are frequent due to the presence of systemic and indirect associations (Krumsiek et al., 2011). Using classical correlation coefficients leads to highly crowded networks where direct and indirect associations are not distinguished (Langfelder and Horvath, 2008). This problem can be successfully overcome using partial correlation (Krumsiek et al., 2011; Valcárcel et al., 2011). In this approach, the correlation between two metabolites is conditioned against the correlation with the remaining metabolites. Consequently, partial correlation allows to discriminate between direct and indirect (i.e., mediated by other metabolites) metabolite correlations. Valcárcel et al. used this approach to build two different networks corresponding to individuals with normal fasting glucose and individuals with prediabetes (Valcárcel et al., 2011). Although few differences were found between individual metabolite concentrations, the network analysis performed in this study revealed significant changes in lipoprotein metabolism, which is known to be associated with diabetes pathophysiology. Netzer et al. used a similar approach to identify highly discriminant metabolites between healthy controls and individuals with obesity (Netzer et al., 2012). In this case, the metabolic network was built using Pearson's correlation coefficient, and the differential metabolites were evaluated by using different network descriptors. In the same study, Netzer et al. used the metabolic differences between two sample groups to build a metabolite ratio network (Netzer et al., 2011). In this approach, the link between two metabolites is scored according to the differences in the ratios between the corresponding metabolites in the two sample groups. The resulting network topology is then based on the metabolic differences between the two studied phenotypes. Recently, Kotze et al. have extended the correlation-based network approach to include prior biological knowledge (Kotze et al., 2013). In this approach, the resulting network is mapped onto known metabolic pathways in order to identify novel links within the metabolic network that may play a key role in the phenotypic trait being studied.

INTEGRATION OF OMICS DATA

Systems biology is the computational modeling of complex biological systems at different molecular levels through the analysis of high-throughput data. Systems biology methods can therefore improve our understanding of the biological processes that are associated with a certain phenotype. These approaches also allow studying how the dysregulation of specific biological pathways is propagated across the biological system. The characterization of the complex and often noisy biological systems has become a major challenge in bioinformatics.

METABOLOMICS INTEGRATION WITH WHOLE GENOME VARIATION

The association between genome-wide genetic variation and high-throughput metabolomic data is one of the current main objectives of omics data integration. The joint analysis of both types of biological data, known as metabolite genome-wide association studies (mGWAS), has allowed the identification of a large number of genomic regions associated with metabolite levels (Gieger et al., 2008; Illig et al., 2010; Suhre et al., 2011a,b; **Table 4**). These

Table 4 | List of studies integrating genomics and metabolomics data.

Cohort size[a]	Metabolites	Biofluid	Metabolomics platform	Objectives	Reference
284	363/40401	Serum	ESI-MS/MS	Study of GIMs	Gieger et al. (2008)
4400	33	Plasma	ESI-MS/MS	Study of GIMs	Hicks et al. (2009)
1809/422	163	Serum	ESI-MS/MS	Study of GIMs	Illig et al. (2010)
1814	163	Serum	ESI-MS/MS	Study of GIMs	Kolz et al. (2009)
862/2031	59	Urine	NMR	Study of GIMs	Suhre et al. (2011b)
1768/1052	276	Serum	UHPLC/MS/MS2, GC/MS	Study of GIMs and overlap with loci of biomedical and pharmaceutical interest	Suhre et al. (2011a)
211	526	Urine and plasma	Multi-platform	Study of GIMs and decomposition of biological population variation in metabolic traits	Nicholson et al. (2011)
4034	153	Plasma	ESI-MS/MS	Study of GIMs and pathway analysis	Demirkan et al. (2012)
8330	216	Serum	NMR	Study of GIMs and heritability of metabolic traits	Kettunen et al. (2012)
6600	130	Serum	NMR	Study of metabolic associations with atherosclerosis using metabolic networks	Inouye et al. (2012)
2076	217	Plasma	HPLC/MS	Study of GIMs and heritability of metabolic traits	Rhee et al. (2013)
7824	486	Plasma	UHPLC/MS/MS2, GC/MS	Study of GIMs, heritability of metabolic traits, and network analysis	Shin et al. (2014)

This table provides an updated list of studies that have integrated metabolomics data with genomics data.

[a] *Studies with discovery and validation cohorts are given as* $N_{discovery}/N_{validation}$.

associations are commonly called genetically influenced metabotypes (GIMs), and could play an important role in the heritability of phenotypic traits. The association between genetic variants and phenotypic traits that often show small association effect sizes can be significantly increased when using intermediate phenotypes like metabolite concentrations (Gieger et al., 2008). These intermediate phenotypes (or endophenotypes) may be characterized by larger effect size associations since they are continuous variables that reflect the actual state of the biological system.

One of the main statistical problems when analyzing the association between genetic variants and metabolite concentrations at a genome-wide level is the large number of tests that must be performed. The number of genetic variants analyzed for each individual by the current high-throughput genotyping technologies usually ranges between 500,000 and $2e^6$. This number of genomic variants can be further increased up to $5–10e^6$ variants with the help of imputation techniques (Howie et al., 2009; Delaneau et al., 2013). Compared to gene-expression data, metabolomic profiles have a much lower number of variables, ranging from 100 s to few 1,000 s. Nevertheless, performing all gene to metabolite association analyses in mGWAS can result in up to $1 \cdot 10^7–1 \cdot 10^{11}$ statistical tests. To date, there are multiple tools that can efficiently perform this large number of quantitative trait analysis like Matrix eQTL (Shabalin, 2012). However, the main limitation of this type of studies is the number of tests that are performed in parallel, and the associated increase in the false positive rate at the nominal ($\alpha = 0.05$) level of significance. Applying a conservative multiple test correction methods like the Bonferroni method leads to extremely high significance thresholds (i.e., corrected α levels $= 1 \cdot 10^{-9}–1 \cdot 10^{-13}$, depending on the total number of performed tests; Gieger et al., 2008; Illig et al., 2010). In order to set a less stringent correction threshold for this type of studies, Demirkan et al. computed the effective number of independent tests by using the number of significant principal components of variation of the metabolomic data (Demirkan et al., 2012). Other studies instead have chosen the genome-wide level of significance commonly used in single-trait GWAS ($\alpha = 5e^{-8}$; McCarthy et al., 2008; Kolz et al., 2009; Tanaka et al., 2009; Rhee et al., 2013).

While most published mGWAS have relied on univariate association tests, there is an increasing effort to develop new multivariate approaches. These approaches have been designed to simultaneously analyze sets of metabolites instead of individual metabolite levels (Klei et al., 2008; Ferreira and Purcell, 2009; O'reilly et al., 2012; Ried et al., 2012; Stephens, 2013). These new approaches have several advantages (Galesloot et al., 2014):

- They take into account the pleiotropic nature of metabolomic data. Subsequently, a genetic variant can be simultaneously associated with multiple metabolites.
- When a genetic variant is associated with multiple metabolites, the aggregated information of the entire set of metabolites increases the statistical power of the analysis (Allison et al., 1998; Zhu and Zhang, 2009).
- By performing a single test for each set of metabolites, the multiple test burden is reduced.

However, one of the main disadvantages of this type of analysis methods is the reduced number of metabolites that can be tested simultaneously. This implies that current metabolite panels (>100 metabolites) cannot be tested together. Inouye et al. overcame this problem by using a two-step design (Inouye et al.,

2012). First, using the metabolite correlation matrix they identified the most relevant metabolic networks using hierarchical clustering. The second step consisted of a multivariate GWAS of each selected network. Each genomic variant was therefore tested a much reduced amount of times and, for each test, the loading of each network metabolite was computed.

Pathway-based approaches are also an important approach for the analysis of genetic variation associated with metabolite levels. As described in Section "Correlation-Based Network Analysis," the methods based on partial correlation coefficients are optimal for the analysis of metabolomic data (Krumsiek et al., 2011). One of these methods, Gaussian Graphical Modeling (GGM), has been recently used to identify unknown metabolites through the integration of metabolomics, GWAS, and pathway data (Krumsiek et al., 2012). Recently, Shin et al. also used GGM to build a complete network of genetic variation associated with human blood metabolite levels (Shin et al., 2014).

METABOLOMICS INTEGRATION WITH OTHER OMICS SCIENCES

Recently, the first study analyzing the association of the genome methylation patterns methylation with metabolic traits has been performed (Petersen et al., 2014). In this study, Petersen et al. used multivariate regression analyses to identify two types of methylome–metabotype associations: (a) associations due to underlying genetic variants and (b) independent associations potentially driven by environmental factors influencing the methylome.

In addition to mGWAS studies, several studies have also explored the association between whole genome gene-expression (i.e., transcriptomics) and metabolomics. The data provided by these two omics sciences have been used, for example, to improve the classification of breast cancers and to explore the correlation between the transcriptional and metabolic levels (Borgan et al., 2010). Borgan et al. used the transcriptional data to classify the breast tumor samples according to previously published tumor types. In a second step, they applied hierarchical clustering on each type of samples using the metabolic data. Using this combined approach, new molecular subtypes of tumors were found. Importantly, these new molecular subtypes were better classified than subtypes based only on gene-expression patterns. Additionally, new biological pathways associated with each molecular subtype could be identified. Using GOrilla software (Eden et al., 2009), they were able to identify potential gene groups regulating each analyzed metabolite. Bjerrum et al. recently combined transcriptomics and metabolomics data from colon biopsies of ulcerative colitis patients. They used O-PLS-DA and multivariate logistic regression models to improve the diagnosis of this autoimmune disease (Bjerrum et al., 2014). Zhang et al. also integrated transcriptomics and metabolomics data to study human pancreatic cancer samples (Zhang et al., 2013). Using a correlation-based network analysis, they identified a set of highly co-regulated and decreased metabolites in these samples and subsequently identified the transcripts correlated with these metabolites.

TOWARD A COMPLETE OMICS INTEGRATION

During the last years, high-throughput technologies have enabled the analysis of the biologic variability at multiple molecular levels. The data obtained from the genome, epigenome, transcriptome, proteome, metabolome, or the microbiome can be now combined using systems biology approaches. However, this group of analytical tools is still in its infancy and major improvements in this field will come in the next years (Chen et al., 2012). 3Omics is one good example of this new type of metabolomic analysis tools. 3Omics is one of the first systems biology tools to provide a full integrative analysis including correlation analysis, co-expression profiling, phenotype mapping, pathway enrichment analysis, and GO enrichment analysis at three molecular levels (transcriptome, proteome, and metabolome; Kuo et al., 2013).

CONCLUSION

Metabolomics is a research field rapidly evolving to allow the fast and accurate analysis of high-throughput data from diverse biological sources. Although the recent methodologies have been able to overcome several challenges of metabolomics data analysis, there is still much room for improvement. In untargeted metabolomic studies, for example, major improvements are still required in automatic metabolite identification and spectral deconvolution. Although a big effort is being done to improve spectral databases, the development of accurate automatic identification algorithms is still subject to the availability of an exhaustive set of reference metabolite spectra.

In addition to the necessary improvements in the analysis workflow, intense efforts are also being done in the standardization of metabolomics data (Salek et al., 2013a). The Metabolomics Standard Initiative (MSI; Fiehn et al., 2007), currently represents the major community effort to define normalization standards in metabolomics. These developments are based on previous high-throughput data standardization initiatives like MIAME in microarray studies (Brazma et al., 2001), and include the use of minimal reported information, common syntax, data format exchange, and common semantics (Field and Sansone, 2006). To date, general guidelines have been proposed (Sumner et al., 2007) that cover relevant areas in metabolomics studies like biological sample processing, analytical technique details (i.e., instrument description, technique-specific acquisition parameters, and sample preparation), instrumental calibration, validation of the quantification method, data pre-processing, metabolite identification, and nomenclature. Very recently, the MetaboLights database (www.ebi.ac.uk/metabolights) has been launched as a repository to archive and distribute data on metabolomics experiments (Steinbeck et al., 2012; Haug et al., 2013; Salek et al., 2013b). Similar to the established public repositories of transcriptomics data (Barrett et al., 2011), the availability of public repositories for metabolomics data will clearly accelerate the progress in this rapidly evolving field.

Omics sciences like metabolomics are increasing our ability to generate knowledge from multiple aspects of biology. In order to achieve these goals, however, the scientific community will require tools and methods that are able to efficiently integrate all the different sources of molecular and phenotypic information. In the near future, increasingly powerful analysis tools will be developed. The access to these methods in an open-source environment will guarantee its dissemination to the largest scientific community possible.

ACKNOWLEDGMENTS

This work was supported by the Spanish Ministry of Economy and Competitiveness grants (IPT-010000-2010-36, PSE-010000-2006-6, and PI12/01362) and by the AGAUR FI grant (2013/00974).

REFERENCES

Adjaye, J., Huntriss, J., Herwig, R., Benkahla, A., Brink, T. C., Wierling, C., et al. (2005). Primary differentiation in the human blastocyst: comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem Cells* 23, 1514–1525. doi:10.1634/stemcells.2005-0113

Aittokallio, T., and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Brief. Bioinformatics* 7, 243–255. doi:10.1093/bib/bbl022

Akiyama, K., Chikayama, E., Yuasa, H., Shimada, Y., Tohge, T., Shinozaki, K., et al. (2008). PRIMe: a web site that assembles tools for metabolomics and transcriptomics. *In silico Biol.* 8, 339–345.

Allison, D. B., Thiel, B., Jean, P. St., Elston, R. C., Infante, M. C., and Schork, N. J. (1998). Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am. J. Hum. Genet.* 63, 1190–1201. doi:10.1086/302038

Alonso, A., Julià, A., Beltran, A., Vinaixa, M., Díaz, M., Ibañez, L., et al. (2011). AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics* 27, 1339–1340. doi:10.1093/bioinformatics/btr138

Alonso, A., Rodríguez, M. A., Vinaixa, M., Tortosa, R., Correig, X., Julià, A., et al. (2013). Focus: a robust workflow for one-dimensional NMR spectral analysis. *Anal. Chem.* 86, 1160–1169. doi:10.1021/ac403110u

Anderson, P., Mahle, D., Doom, T., Reo, N., Delraso, N., and Raymer, M. (2011). Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data. *Metabolomics* 7, 179–190. doi:10.1007/s11306-010-0242-7

Anderson, P., Reo, N., Delraso, N., Doom, T., and Raymer, M. (2008). Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics. *Metabolomics* 4, 261–272. doi:10.1007/s11306-008-0117-3

Armitage, E. G., and Barbas, C. (2014). Metabolomics in cancer biomarker discovery: current trends and future perspectives. *J. Pharm. Biomed. Anal.* 87, 1–11. doi:10.1016/j.jpba.2013.08.041

Astle, W., De Iorio, M., Richardson, S., Stephens, D., and Ebbels, T. (2012). A Bayesian model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures. *J. Am. Stat. Assoc.* 107, 1259–1271. doi:10.1093/bioinformatics/bts308

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., et al. (2011). NCBI GEO: archive for functional genomics data sets – 10 years on. *Nucleic Acids Res.* 39, D1005–D1010. doi:10.1093/nar/gkq1184

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Soc. Series B Stat. Methodol.* 57, 289–300.

Bingol, K., Bruschweiler-Li, L., Li, D.-W., and Brüschweiler, R. (2014). Customized metabolomics database for the analysis of NMR 1H – 1H TOCSY and 13C – 1H HSQC-TOCSY spectra of complex mixtures. *Anal. Chem.* 86, 5494–5501. doi:10.1021/ac500979g

Bingol, K., Li, D.-W., Bruschweiler-Li, L., Cabrera, O., Megraw, T., Zhang, F., et al. (2015). Unified and isomer-specific NMR metabolomics database for the accurate analysis of 13C-1H HSQC spectra. *ACS Chem. Biol.* 10, 452–459. doi:10.1021/cb5006382

Bingol, K., Zhang, F., Bruschweiler-Li, L., and Brüschweiler, R. (2012). TOCCATA: a customized carbon total correlation spectroscopy NMR metabolomics database. *Anal. Chem.* 84, 9395–9401. doi:10.1021/ac302197e

Biswas, A., Mynampati, K. C., Umashankar, S., Reuben, S., Parab, G., Rao, R., et al. (2010). MetDAT: a modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation. *Bioinformatics* 26, 2639–2640. doi:10.1093/bioinformatics/btq436

Bjerrum, J., Rantalainen, M., Wang, Y., Olsen, J., and Nielsen, O. (2014). Integration of transcriptomics and metabonomics: improving diagnostics, biomarker identification and phenotyping in ulcerative colitis. *Metabolomics* 10, 280–290. doi:10.1007/s11306-013-0580-3

Blümich, B., and Callaghan, P. T. (1995). Principles of nuclear magnetic resonance microscopy. Oxford University Press, Oxford, 1993, 492 pp, £25. ISBN 0 198 53997 5. *Magn. Reson. Chem.* 33, 322–322. doi:10.1002/mrc.1260330417

Borgan, E., Sitter, B., Lingjaerde, O., Johnsen, H., Lundgren, S., Bathen, T., et al. (2010). Merging transcriptomics and metabolomics – advances in breast cancer profiling. *BMC Cancer* 10:628. doi:10.1186/1471-2407-10-628

Bothwell, J. H., and Griffin, J. L. (2011). An introduction to biological nuclear magnetic resonance spectroscopy. *Biol. Rev. Camb. Philos. Soc.* 86, 493–510. doi:10.1111/j.1469-185X.2010.00157.x

Bouatra, S., Aziat, F., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., et al. (2013). The human urine metabolome. *PLoS ONE* 8:e73076. doi:10.1371/journal.pone.0073076

Brauer, M. J., Yuan, J., Bennett, B. D., Lu, W., Kimball, E., Botstein, D., et al. (2006). Conservation of the metabolomic response to starvation across two divergent microbes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 19302–19307. doi:10.1073/pnas.0609508103

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365–371. doi:10.1038/ng1201-365

Bro, R., and Smilde, A. K. (2014). Principal component analysis. *Anal. Methods* 6, 2812–2831. doi:10.1039/c3ay41907j

Broadhurst, D., and Kell, D. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2, 171–196. doi:10.1007/s11306-006-0037-z

Burton, L., Ivosev, G., Tate, S., Impey, G., Wingate, J., and Bonner, R. (2008). Instrumental and experimental effects in LC – MS-based metabolomics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 871, 227–235. doi:10.1016/j.jchromb.2008.04.044

Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., and Trygg, J. (2006). OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J. Chemom.* 20, 341–351. doi:10.1002/cem.1006

Camacho, D., De La Fuente, A., and Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics* 1, 53–63. doi:10.1007/s11306-005-1107-3

Carroll, A., Badger, M., and Harvey Millar, A. (2010). The Metabolome Express Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinformatics* 11:376. doi:10.1186/1471-2105-11-376

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., et al. (2008). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 36, D623–D631. doi:10.1093/nar/gkm900

Chadeau-Hyam, M., Ebbels, T. M. D., Brown, I. J., Chan, Q., Stamler, J., Huang, C. C., et al. (2010). Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. *J. Proteome Res.* 9, 4620–4627. doi:10.1021/pr1003449

Chae, L., Kim, T., Nilo-Poyanco, R., and Rhee, S. Y. (2014). Genomic signatures of specialized metabolism in plants. *Science* 344, 510–513. doi:10.1126/science.1252076

Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L., Lam, H. Y., Chen, R., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148, 1293–1307. doi:10.1016/j.cell.2012.02.009

Chylla, R. A., Hu, K., Ellinger, J. J., and Markley, J. L. (2011). Deconvolution of two-dimensional NMR spectra by fast maximum likelihood reconstruction: application to quantitative metabolomics. *Anal. Chem.* 83, 4871–4880. doi:10.1021/ac200536b

Clifford, D., Stone, G., Montoliu, I., Rezzi, S., Martin, F.-P., Guy, P., et al. (2009). Alignment using variable penalty dynamic time warping. *Anal. Chem.* 81, 1000–1007. doi:10.1021/ac802041e

Cottret, L., Wildridge, D., Vinson, F., Barrett, M. P., Charles, H., Sagot, M.-F., et al. (2010). MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res.* 38, W132–W137. doi:10.1093/nar/gkq312

Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K., and Lindon, J. C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal. Chem.* 78, 2262–2267. doi:10.1021/ac0519312

Cui, Q., Lewis, I. A., Hegeman, A. D., Anderson, M. E., Li, J., Schulte, C. F., et al. (2008). Metabolite identification via the Madison metabolomics consortium database. *Nat. Biotechnol.* 26, 162–164. doi:10.1038/nbt0208-162

De Meyer, T., Sinnaeve, D., Van Gasse, B., Tsiporkova, E., Rietzschel, E. R., De Buyzere, M. L., et al. (2008). NMR-based characterization of metabolic alterations in

hypertension using an adaptive, intelligent binning algorithm. *Anal. Chem.* 80, 3783–3790. doi:10.1021/ac7025964

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* 10, 5–6. doi:10.1038/nmeth.2307

Demirkan, A., Van Duijn, C. M., Ugocsai, P., Isaacs, A., Pramstaller, P. P., Liebisch, G., et al. (2012). Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet.* 8:e1002490. doi:10.1371/journal.pgen.1002490

Dietrich, W., Rüdel, C. H., and Neumann, M. (1991). Fast and precise automatic baseline correction of one- and two-dimensional NMR spectra. *J. Magn. Reson. (1969)* 91, 1–11. doi:10.1016/0022-2364(91)90402-F

Du, P., Kibbe, W. A., and Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* 22, 2059–2065. doi:10.1093/bioinformatics/btl355

Du, X., and Zeisel, S. H. (2013). Spectral deconvolution for gas chromatography mass spectrometry-based metabolomics: current status and future perspectives. *Comput. Struct. Biotechnol. J.* 4, e201301013. doi:10.5936/csbj.201301013

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48. doi:10.1186/1471-2105-10-48

Eilers, P. H. C. (2003). Parametric time warping. *Anal. Chem.* 76, 404–411. doi:10.1021/ac034800e

El-Aneed, A., Cohen, A., and Banoub, J. (2009). Mass spectrometry, review of the basics: electrospray, MALDI, and commonly used mass analyzers. *Appl. Spectrosc. Rev.* 44, 210–230. doi:10.1080/05704920902717872

Ellinger, J. J., Chylla, R. A., Ulrich, E. L., and Markley, J. L. (2013). Databases and software for NMR-based metabolomics. *Curr. Metabolomics* 1, 28–40. doi:10.2174/2213235X11301010028

Fernández-Albert, F., Llorach, R., Andrés-Lacueva, C., and Perera, A. (2014). An R package to analyse LC/MS metabolomic data: MAIT (metabolite automatic identification toolkit). *Bioinformatics* 30, 1937–1939. doi:10.1093/bioinformatics/btu136

Ferreira, M. A., and Purcell, S. M. (2009). A multivariate test of association. *Bioinformatics* 25, 132–133. doi:10.1093/bioinformatics/btn563

Fiehn, O., Robertson, D., Griffin, J., Van Der Werf, M., Nikolau, B., Morrison, N., et al. (2007). The metabolomics standards initiative (MSI). *Metabolomics* 3, 175–178. doi:10.1007/s11306-007-0070-6

Field, D., and Sansone, S.-A. (2006). A special issue on data standards. *OMICS* 10, 84–93. doi:10.1089/omi.2006.10.84

Fonville, J. M., Richards, S. E., Barton, R. H., Boulange, C. L., Ebbels, T. M. D., Nicholson, J. K., et al. (2010). The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *J. Chemom.* 24, 636–649. doi:10.1002/cem.1359

Forshed, J., Schuppe-Koistinen, I., and Jacobsson, S. P. (2003). Peak alignment of NMR signals by means of a genetic algorithm. *Anal. Chim. Acta* 487, 189–199. doi:10.1016/S0003-2670(03)00570-1

Fuhrer, T., and Zamboni, N. (2015). High-throughput discovery metabolomics. *Curr. Opin. Biotechnol.* 31, 73–78. doi:10.1016/j.copbio.2014.08.006

Fukushima, A., and Kusano, M. (2013). Recent progress in the development of metabolome databases for plant systems biology. *Front. Plant Sci.* 4:73. doi:10.3389/fpls.2013.00073

Galesloot, T. E., Van Steen, K., Kiemeney, L. A., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PLoS ONE* 9:e95923. doi:10.1371/journal.pone.0095923

Gao, J., Tarcea, V. G., Karnovsky, A., Mirel, B. R., Weymouth, T. E., Beecher, C. W., et al. (2010). MetScape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics* 26, 971–973. doi:10.1093/bioinformatics/btq048

García-Alcalde, F., García-López, F., Dopazo, J., and Conesa, A. (2011). Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics* 27, 137–139. doi:10.1093/bioinformatics/btq594

Gaude, E., Chignola, F., Spiliotopoulos, D., Spitaleri, A., Ghitti, M., García-Manteiga, J. M., et al. (2013). muma, An R package for metabolomics univariate and multivariate statistical analysis. *Curr. Metabolomics* 1, 180–189. doi:10.2174/2213235X11301020005

Gibbons, H., O'gorman, A., and Brennan, L. (2015). Metabolomics as a tool in nutritional research. *Curr. Opin. Lipidol.* 26, 30–34. doi:10.1097/MOL.0000000000000140

Gieger, C., Geistlinger, L., Altmaier, E., Hrabé De Angelis, M., Kronenberg, F., Meitinger, T., et al. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* 4:e1000282. doi:10.1371/journal.pgen.1000282

Gika, H. G., Theodoridis, G. A., Plumb, R. S., and Wilson, I. D. (2014). Current practice of liquid chromatography – mass spectrometry in metabolomics and metabonomics. *J. Pharm. Biomed. Anal.* 87, 12–25. doi:10.1016/j.jpba.2013.06.032

Gika, H. G., Theodoridis, G. A., and Wilson, I. D. (2008). Liquid chromatography and ultra-performance liquid chromatography – mass spectrometry fingerprinting of human urine: sample stability under different handling and storage conditions for metabonomics studies. *J. Chromatogr. A* 1189, 314–322. doi:10.1016/j.chroma.2007.10.066

Giskeødegård, G. F., Bloemberg, T. G., Postma, G., Sitter, B., Tessem, M.-B., Gribbestad, I. S., et al. (2010). Alignment of high resolution magic angle spinning magnetic resonance spectra using warping methods. *Anal. Chim. Acta* 683, 1–11. doi:10.1016/j.aca.2010.09.026

Goeman, J. J., Van De Geer, S. A., De Kort, F., and Van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20, 93–99. doi:10.1093/bioinformatics/btg382

Goodwin, C. R., Sherrod, S. D., Marasco, C. C., Bachmann, B. O., Schramm-Sapyta, N., Wikswo, J. P., et al. (2014). Phenotypic mapping of metabolic profiles using self-organizing maps of high-dimensional mass spectrometry data. *Anal. Chem.* 86, 6563–6571. doi:10.1021/ac5010794

Hao, J., Astle, W., De Iorio, M., and Ebbels, T. M. D. (2012). BATMAN – an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* 28, 2088–2090. doi:10.1093/bioinformatics/bts308

Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J. G., and Ebbels, T. M. D. (2014). Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat. Protoc.* 9, 1416–1427. doi:10.1038/nprot.2014.090

Haug, K., Salek, R. M., Conesa, P., Hastings, J., De Matos, P., Rijnbeek, M., et al. (2013). MetaboLights – an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 41, D781–D786. doi:10.1093/nar/gks1004

Hicks, A. A., Pramstaller, P. P., Johansson, A., Vitart, V., Rudan, I., Ugocsai, P., et al. (2009). Genetic determinants of circulating sphingolipid concentrations in European populations. *PLoS Genet.* 5:e1000672. doi:10.1371/journal.pgen.1000672

Hiller, K., Hangebrauk, J., Jäger, C., Spura, J., Schreiber, K., and Schomburg, D. (2009). MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Anal. Chem.* 81, 3429–3439. doi:10.1021/ac802689c

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., et al. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* 45, 703–714. doi:10.1002/jms.1777

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi:10.1371/journal.pgen.1000529

Hummel, J., Selbig, J., Walther, D., and Kopka, J. (2007). "The Golm Metabolome Database: a database for GC-MS based metabolite profiling," in *Metabolomics*, eds J. Nielsen and M. Jewett (Berlin: Springer), 75–95.

Hummel, M., Meister, R., and Mansmann, U. (2008). GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics* 24, 78–85. doi:10.1093/bioinformatics/btm531

Illig, T., Gieger, C., Zhai, G., Romisch-Margl, W., Wang-Sattler, R., Prehn, C., et al. (2010). A genome-wide perspective of genetic variation in human metabolism. *Nat. Genet.* 42, 137–141. doi:10.1038/ng.507

Inouye, M., Ripatti, S., Kettunen, J., Lyytikäinen, L.-P., Oksala, N., Laurila, P.-P., et al. (2012). Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genet.* 8:e1002907. doi:10.1371/journal.pgen.1002907

Jacob, D., Deborde, C., and Moing, A. (2013). An efficient spectra processing method for metabolite identification from 1H-NMR metabolomics data. *Anal. Bioanal. Chem.* 405, 5049–5061. doi:10.1007/s00216-013-6852-y

Jewison, T., Su, Y., Disfany, F. M., Liang, Y., Knox, C., Maciejewski, A., et al. (2014). SMPDB 2.0: big improvements to the small molecule pathway database. *Nucleic Acids Res.* 42, D478–D484. doi:10.1093/nar/gkt1067

Jiang, W., Zhang, Z.-M., Yun, Y., Zhan, D.-J., Zheng, Y.-B., Liang, Y.-Z., et al. (2013). Comparisons of five algorithms for chromatogram alignment. *Chromatographia* 76, 1067–1078. doi:10.1007/s10337-013-2513-8

Julià, A., Alonso, A., and Marsal, S. (2014). Metabolomics in rheumatic diseases. *Int. J. Clin. Rheumatol.* 9, 353–369. doi:10.2217/ijr.14.25

Jung, S.-H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics* 21, 3097–3104. doi:10.1093/bioinformatics/bti456

Kaddurah-Daouk, R., Kristal, B. S., and Weinshilboum, R. M. (2008). Metabolomics: a global biochemical approach to drug response and disease. *Annu. Rev. Pharmacol. Toxicol.* 48, 653–683. doi:10.1146/annurev.pharmtox.48.113006.094715

Kamburov, A., Cavill, R., Ebbels, T. M. D., Herwig, R., and Keun, H. C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* 27, 2917–2918. doi:10.1093/bioinformatics/btr499

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi:10.1093/nar/gkr988

Karakach, T. K., Knight, R., Lenz, E. M., Viant, M. R., and Walter, J. A. (2009). Analysis of time course 1H NMR metabolomics data by multivariate curve resolution. *Magn. Reson. Chem.* 47, S105–S117. doi:10.1002/mrc.2535

Karnovsky, A., Weymouth, T., Hull, T., Tarcea, V. G., Scardoni, G., Laudanna, C., et al. (2012). MetScape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* 28, 373–380. doi:10.1093/bioinformatics/btr661

Kelder, T., Van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., et al. (2012). WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* 40, D1301–D1307. doi:10.1093/nar/gkr1074

Kell, D. B., and Goodacre, R. (2014). Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. *Drug Discov. Today* 19, 171–182. doi:10.1016/j.drudis.2013.07.014

Kemsley, E. K. (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometr. Intell. Lab. Syst.* 33, 47–61. doi:10.1186/1471-2105-10-213

Kettunen, J., Tukiainen, T., Sarin, A.-P., Ortega-Alonso, A., Tikkanen, E., Lyytikainen, L.-P., et al. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* 44, 269–276. doi:10.1038/ng.1073

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8:e1002375. doi:10.1371/journal.pcbi.1002375

Kim, Y., Koo, I., Jung, B. H., Chung, B. C., and Lee, D. (2010). Multivariate classification of urine metabolome profiles for breast cancer diagnosis. *BMC Bioinformatics* 11:S4. doi:10.1186/1471-2105-11-S2-S4

Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet. Epidemiol.* 32, 9–19. doi:10.1002/gepi.20257

Kohl, S., Klein, M., Hochrein, J., Oefner, P., Spang, R., and Gronwald, W. (2012). State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics* 8, 146–160. doi:10.1007/s11306-011-0350-z

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., et al. (2000). Self organization of a massive document collection. *IEEE Trans. Neural Netw.* 11, 574–585. doi:10.1109/72.846729

Kolz, M., Johnson, T., Sanna, S., Teumer, A., Vitart, V., Perola, M., et al. (2009). Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genet.* 5:e1000504. doi:10.1371/journal.pgen.1000504

Kotze, H., Armitage, E., Sharkey, K., Allwood, J., Dunn, W., Williams, K., et al. (2013). A novel untargeted metabolomics correlation-based network analysis incorporating human metabolic reconstructions. *BMC Syst. Biol.* 7:107. doi:10.1186/1752-0509-7-107

Krumsiek, J., Suhre, K., Evans, A. M., Mitchell, M. W., Mohney, R. P., Milburn, M. V., et al. (2012). Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet.* 8:e1003005. doi:10.1371/journal.pgen.1003005

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* 5:21. doi:10.1186/1752-0509-5-21

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., and Neumann, S. (2011). CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* 84, 283–289. doi:10.1021/ac202450g

Kühn, C. (2012). "Metabolomics in animal breeding," in *Genetics Meets Metabolomics*, ed. K. Suhre (New York, NY: Springer), 107–123.

Kuo, T.-C., Tian, T.-F., and Tseng, Y. (2013). 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst. Biol.* 7:64. doi:10.1186/1752-0509-7-64

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi:10.1186/1471-2105-9-559

Lee, G.-C., and Woodruff, D. L. (2004). Beam search for peak alignment of NMR signals. *Anal. Chim. Acta* 513, 413–416. doi:10.1016/j.aca.2004.02.068

Lewis, I. A., Schommer, S. C., and Markley, J. L. (2009). rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn. Reson. Chem.* 47, S123–S126. doi:10.1002/mrc.2526

Lommen, A., and Kools, H. (2012). MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics* 8, 719–726. doi:10.1007/s11306-011-0369-1

Ludwig, C., Easton, J., Lodi, A., Tiziani, S., Manzoor, S., Southam, A., et al. (2012). Birmingham metabolite library: a publicly accessible database of 1-D 1H and 2-D 1H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics* 8, 8–18. doi:10.1007/s11306-011-0347-7

Ludwig, C., and Gunther, U. (2011). MetaboLab – advanced NMR data processing and analysis for metabolomics. *BMC Bioinformatics* 12:366. doi:10.1186/1471-2105-12-366

Luts, J., Ojeda, F., Van De Plas, R., De Moor, B., Van Huffel, S., and Suykens, J. A. (2010). A tutorial on support vector machine-based methods for classification problems in chemometrics. *Anal. Chim. Acta* 665, 129–145. doi:10.1016/j.aca.2010.03.030

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., et al. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* 3, 135. doi:10.1038/msb4100177

Madsen, R., Lundstedt, T., and Trygg, J. (2010). Chemometrics in metabolomics – A review in human disease diagnosis. *Anal. Chim. Acta* 659, 23–33. doi:10.1016/j.aca.2009.11.042

Mahadevan, S., Shah, S. L., Marrie, T. J., and Slupsky, C. M. (2008). Analysis of metabolomic data using support vector machines. *Anal. Chem.* 80, 7562–7570. doi:10.1021/ac800954c

Mäkinen, V.-P., Soininen, P., Forsblom, C., Parkkonen, M., Ingman, P., Kaski, K., et al. (2008). 1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. *Mol. Syst. Biol.* 4, 167. doi:10.1038/msb4100205

Mamas, M., Dunn, W., Neyses, L., and Goodacre, R. (2011). The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Arch. Toxicol.* 85, 5–17. doi:10.1007/s00204-010-0609-6

Marion, D. (2013). An introduction to biological NMR spectroscopy. *Mol. Cell Proteomics* 12, 3006–3025. doi:10.1074/mcp.O113.030239

Martin, F. O.-P. J., Montoliu, I., Kochhar, S., and Rezzi, S. (2010). Chemometric strategy for modeling metabolic biological space along the gastrointestinal tract and assessing microbial influences. *Anal. Chem.* 82, 9803–9811. doi:10.1021/ac102015n

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369. doi:10.1038/nrg2344

Meinicke, P., Lingner, T., Kaever, A., Feussner, K., Göbel, C., Feussner, I., et al. (2008). Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms Mol. Biol.* 3, 1–18. doi:10.1186/1748-7188-3-9

Melamud, E., Vastag, L., and Rabinowitz, J. D. (2010). Metabolomic analysis and visualization engine for LC-MS data. *Anal. Chem.* 82, 9818–9826. doi:10.1021/ac1021166

Mercier, P., Lewis, M., Chang, D., Baker, D., and Wishart, D. (2011). Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra. *J. Biomol. NMR* 49, 307–323. doi:10.1007/s10858-011-9480-x

Meyer, U. A., Zanger, U. M., and Schwab, M. (2013). Omics and drug response. *Annu. Rev. Pharmacol. Toxicol.* 53, 475–502. doi:10.1146/annurev-pharmtox-010510-100502

Montoliu, I., Martin, F.-P. J., Collino, S., Rezzi, S., and Kochhar, S. (2009). Multivariate modeling strategy for intercompartmental analysis of tissue and plasma 1H NMR spectrotypes. *J. Proteome Res.* 8, 2397–2406. doi:10.1021/pr8010205

Netzer, M., Kugler, K. G., Müller, L. A., Weinberger, K. M., Graber, A., Baumgartner, C., et al. (2012). A network-based feature selection approach to identify metabolic signatures in disease. *J. Theor. Biol.* 310, 216–222. doi:10.1016/j.jtbi.2012.06.003

Netzer, M., Weinberger, K., Handler, M., Seger, M., Fang, X., Kugler, K., et al. (2011). Profiling the human response to physical exercise: a computational strategy for the identification and kinetic analysis of metabolic biomarkers. *J. Clin. Bioinforma.* 1, 34. doi:10.1186/2043-9113-1-34

Ni, Y., Qiu, Y., Jiang, W., Suttlemyre, K., Su, M., Zhang, W., et al. (2012). ADAP-GC 2.0: deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies. *Anal. Chem.* 84, 6619–6629. doi:10.1021/ac300898h

Nicholson, G., Rantalainen, M., Li, J. V., Maher, A. D., Malmodin, D., Ahmadi, K. R., et al. (2011). A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet.* 7:e1002270. doi:10.1371/journal.pgen.1002270

Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., et al. (2012). Host-gut microbiota metabolic interactions. *Science* 336, 1262–1267. doi:10.1126/science.1223813

Niu, W., Knight, E., Xia, Q., and McGarvey, B. D. (2014). Comparative evaluation of eight software programs for alignment of gas chromatography – mass spectrometry chromatograms in metabolomics experiments. *J. Chromatogr. A* 1374, 199–206. doi:10.1016/j.chroma.2014.11.005

O'reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, M.-R., et al. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE* 7:e34861. doi:10.1371/journal.pone.0034861

Orešič, M. (2009). Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction. *Nutr. Metab. Cardiovasc. Dis.* 19, 816–824. doi:10.1016/j.numecd.2009.04.018

Patti, G. J., Yanes, O., and Siuzdak, G. (2012). Innovation: metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* 13, 263–269. doi:10.1038/nrm3314

Peré-Trepat, E., Ginebreda, A., and Tauler, R. (2007). Comparison of different multiway methods for the analysis of geographical metal distributions in fish, sediments and river waters in Catalonia. *Chemometr. Intell. Lab. Syst.* 88, 69–83. doi:10.1016/j.chemolab.2006.09.009

Petersen, A.-K., Zeilinger, S., Kastenmüller, G., Römisch-Margl, W., Brugger, M., Peters, A., et al. (2014). Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. *Hum. Mol. Genet.* 23, 534–545. doi:10.1093/hmg/ddt430

Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11:395. doi:10.1186/1471-2105-11-395

Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., et al. (2011). The human serum metabolome. *PLoS ONE* 6:e16957. doi:10.1371/journal.pone.0016957

Putri, S. P., Yamamoto, S., Tsugawa, H., and Fukusaki, E. (2013). Current metabolomics: technological advances. *J. Biosci. Bioeng.* 116, 9–16. doi:10.1016/j.jbiosc.2013.01.004

Qi, X., and Zhang, D. (2014). Plant metabolomics and metabolic biology. *J. Integr. Plant Biol.* 56, 814–815. doi:10.1111/jipb.12247

Rafiei, A., and Sleno, L. (2015). Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis. *Rapid Commun. Mass Spectrom.* 29, 119–127. doi:10.1002/rcm.7094

Rasmussen, L., Savorani, F., Larsen, T., Dragsted, L., Astrup, A., and Engelsen, S. (2011). Standardization of factors that influence human urine metabolomics. *Metabolomics* 7, 71–83. doi:10.1007/s11306-010-0234-7

Ravanbakhsh, S., Liu, P., Mandal, R., Grant, J. R., Wilson, M., Eisner, R., et al. (2014). Accurate, fully-automated NMR spectral profiling for metabolomics. *arXiv* 1409–1456.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19, 368–375. doi:10.1093/bioinformatics/btf877

Reo, N. V. (2002). NMR-based metabolomics. *Drug Chem. Toxicol.* 25, 375–382. doi:10.1081/DCT-120014789

Rhee, E. P., Ho, J. E., Chen, M. H., Shen, D., Cheng, S., Larson, M. G., et al. (2013). A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab.* 18, 130–143. doi:10.1016/j.cmet.2013.06.013

Ried, J. S., Döring, A., Oexle, K., Meisinger, C., Winkelmann, J., Klopp, N., et al. (2012). PSEA: phenotype set enrichment analysis – a new method for analysis of multiple phenotypes. *Genet. Epidemiol.* 36, 244–252. doi:10.1002/gepi.21617

Roberts, L. D., Souza, A. L., Gerszten, R. E., and Clish, C. B. (2012). Targeted metabolomics. *Curr. Protoc. Mol. Biol.* Chapter 30, 1–24. doi:10.1002/0471142727.mb3002s98

Robertson, D. G., and Frevert, U. (2013). Metabolomics in drug discovery and development. *Clin. Pharmacol. Ther.* 94, 559–561. doi:10.1038/clpt.2013.120

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi:10.1186/1471-2105-12-77

Robinette, S. L., Zhang, F., Brüschweiler-Li, L., and Brüschweiler, R. (2008). Web server based complex mixture analysis by NMR. *Anal. Chem.* 80, 3606–3611. doi:10.1021/ac702530t

Rohn, H., Junker, A., Hartmann, A., Grafahrend-Belau, E., Treutler, H., Klapperstuck, M., et al. (2012). VANTED v2: a framework for systems biology applications. *BMC Syst. Biol.* 6:139. doi:10.1186/1752-0509-6-139

Sakurai, T., Yamada, Y., Sawada, Y., Matsuda, F., Akiyama, K., Shinozaki, K., et al. (2013). PRIMe update: innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. *Plant Cell Physiol.* 54, e5–e5. doi:10.1093/pcp/pcs184

Salek, R., Steinbeck, C., Viant, M., Goodacre, R., and Dunn, W. (2013a). The role of reporting standards for metabolite annotation and identification in metabolomic studies. *Gigascience* 2, 13. doi:10.1186/2047-217X-2-13

Salek, R. M., Haug, K., Conesa, P., Hastings, J., Williams, M., Mahendraker, T., et al. (2013b). The MetaboLights repository: curation challenges in metabolomics. *Database (Oxford)* 2013:bat029. doi:10.1093/database/bat029

Savorani, F., Tomasi, G., and Engelsen, S. B. (2010). Icoshift: a versatile tool for the rapid alignment of 1D NMR spectra. *J. Magn. Reson.* 202, 190–202. doi:10.1016/j.jmr.2009.11.012

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353–1358. doi:10.1093/bioinformatics/bts163

Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., et al. (2014). An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46, 543–550. doi:10.1038/ng.2982

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* 21, 3940–3941. doi:10.1093/bioinformatics/bti623

Smith, C. A., Want, E. J., O'maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78, 779–787. doi:10.1021/ac051437y

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi:10.1093/bioinformatics/btq675

Sousa, S. A. A., Magalhães, A., and Ferreira, M. M. C. (2013). Optimized bucketing for NMR spectra: three case studies. *Chemometr. Intell. Lab. Syst.* 122, 93–102. doi:10.1016/j.chemolab.2013.01.006

Sreekumar, A., Poisson, L. M., Rajendiran, T. M., Khan, A. P., Cao, Q., Yu, J., et al. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* 457, 910–914. doi:10.1038/nature07762

Stein, S. E. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.* 10, 770–781. doi:10.1016/S1044-0305(99)00047-1

Steinbeck, C., Conesa, P., Haug, K., Mahendraker, T., Williams, M., Maguire, E., et al. (2012). MetaboLights: towards a new COSMOS of metabolomics data management. *Metabolomics* 8, 757–760. doi:10.1007/s11306-012-0462-0

Steinbeck, C., Krause, S., and Kuhn, S. (2003). NMRShiftDB constructing a free chemical information system with open-source components. *J. Chem. Inf. Comput. Sci.* 43, 1733–1739.

Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS ONE* 8:e65245. doi:10.1371/journal.pone.0065245

Steuer, R. (2006). Review: on the analysis and interpretation of correlations in metabolomic data. *Brief. Bioinformatics* 7, 151–158. doi:10.1093/bib/bbl009

Steuer, R., Kurths, J., Fiehn, O., and Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics* 19, 1019–1026. doi:10.1093/bioinformatics/btg120

Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., Hussong, R., Lange, E., et al. (2008). OpenMS – an open-source software framework for mass spectrometry. *BMC Bioinformatics* 9:163. doi:10.1186/1471-2105-9-163

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E. A., Glass, C. K., et al. (2007). LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* 35, D527–D532. doi:10.1093/nar/gkl838

Suhre, K., Shin, S.-Y., Petersen, A.-K., Mohney, R. P., Meredith, D., Wagele, B., et al. (2011a). Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 477, 54–60. doi:10.1038/nature10354

Suhre, K., Wallaschofski, H., Raffler, J., Friedrich, N., Haring, R., Michael, K., et al. (2011b). A genome-wide association study of metabolic traits in human urine. *Nat. Genet.* 43, 565–569. doi:10.1038/ng.837

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., et al. (2007). Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI). *Metabolomics* 3, 211–221. doi:10.1007/s11306-007-0082-2

Szymanska, E., Saccenti, E., Smilde, A., and Westerhuis, J. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* 8, 3–16. doi:10.1007/s11306-011-0330-3

Tanaka, T., Shen, J., Abecasis, G. R., Kisialiou, A., Ordovas, J. M., Guralnik, J. M., et al. (2009). Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI study. *PLoS Genet.* 5:e1000338. doi:10.1371/journal.pgen.1000338

Tapp, H. S., and Kemsley, E. K. (2009). Notes on the practical utility of OPLS. *Trends Analyt. Chem.* 28, 1322–1327. doi:10.1016/j.trac.2009.08.006

Tautenhahn, R., Bottcher, C., and Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* 9:504. doi:10.1186/1471-2105-9-504

Tautenhahn, R., Cho, K., Uritboonthai, W., Zhu, Z., Patti, G. J., and Siuzdak, G. (2012a). An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat. Biotechnol.* 30, 826–828. doi:10.1038/nbt.2348

Tautenhahn, R., Patti, G. J., Rinehart, D., and Siuzdak, G. (2012b). XCMS online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* 84, 5035–5039. doi:10.1021/ac300698c

Tautenhahn, R., Patti, G. J., Kalisiak, E., Miyamoto, T., Schmidt, M., Lo, F. Y., et al. (2010). metaXCMS: second-order analysis of untargeted metabolomics data. *Anal. Chem.* 83, 696–700. doi:10.1021/ac102980g

Theodoridis, G., Gika, H. G., and Wilson, I. D. (2011). Mass spectrometry-based holistic analytical approaches for metabolite profiling in systems biology studies. *Mass Spectrom. Rev.* 30, 884–906. doi:10.1002/mas.20306

Tomasi, G., Van Den Berg, F., and Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemom* 18, 231–241. doi:10.1002/cem.859

Townsend, M. K., Clish, C. B., Kraft, P., Wu, C., Souza, A. L., Deik, A. A., et al. (2013). Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. *Clin. Chem.* 59, 1657–1667. doi:10.1373/clinchem.2012.199133

Trygg, J., and Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *J. Chemom.* 16, 119–128. doi:10.1002/cem.695

Tulpan, D., Leger, S., Belliveau, L., Culf, A., and Cuperlovic-Culf, M. (2011). MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC Bioinformatics* 12:400. doi:10.1186/1471-2105-12-400

Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., et al. (2008). BioMagResBank. *Nucleic Acids Res.* 36, D402–D408. doi:10.1093/nar/gkm957

Valcárcel, B., Würtz, P., Seich Al Basatena, N.-K., Tukiainen, T., Kangas, A. J., Soininen, P., et al. (2011). A differential network approach to exploring differences between biological states: an application to prediabetes. *PLoS ONE* 6:e24702. doi:10.1371/journal.pone.0024702

Van Den Oord, E. J. C. G. (2008). Controlling false discoveries in genetic studies. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 147B, 637–644. doi:10.1002/ajmg.b.30650

Van Nederkassel, A. M., Daszykowski, M., Eilers, P. H. C., and Heyden, Y. V. (2006). A comparison of three algorithms for chromatograms alignment. *J. Chromatogr. A* 1118, 199–210. doi:10.1016/j.chroma.2006.03.114

Veselkov, K. A., Lindon, J. C., Ebbels, T. M. D., Crockford, D., Volynkin, V. V., Holmes, E., et al. (2008). Recursive segment-wise peak alignment of biological 1H NMR spectra for improved metabolic biomarker recovery. *Anal. Chem.* 81, 56–66. doi:10.1021/ac8011544

Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J. J., and Yanes, O. (2012). A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites* 2, 775–795. doi:10.3390/metabo2040775

Vu, T., and Laukens, K. (2013). Getting your peaks in line: a review of alignment methods for NMR spectral data. *Metabolites* 3, 259–276. doi:10.3390/metabo3020259

Vu, T., Valkenborg, D., Smets, K., Verwaest, K., Dommisse, R., Lemiere, F., et al. (2011). An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics* 12:405. doi:10.1186/1471-2105-12-405

Wang, T., Shao, K., Chu, Q., Ren, Y., Mu, Y., Qu, L., et al. (2009). Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC Bioinformatics* 10:83. doi:10.1186/1471-2105-10-83

Ward, J. L., Baker, J. M., and Beale, M. H. (2007). Recent applications of NMR spectroscopy in plant metabolomics. *FEBS J.* 274, 1126–1131. doi:10.1111/j.1742-4658.2007.05675.x

Weljie, A. M., Newton, J., Mercier, P., Carlson, E., and Slupsky, C. M. (2006). Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal. Chem.* 78, 4430–4442. doi:10.1021/ac060209g

Westerhuis, J., Hoefsloot, H. J., Smit, S., Vis, D., Smilde, A., Van Velzen, E. J., et al. (2008). Assessment of PLSDA cross validation. *Metabolomics* 4, 81–89. doi:10.1007/s11306-007-0099-6

Wikoff, W. R., Anfora, A. T., Liu, J., Schultz, P. G., Lesley, S. A., Peters, E. C., et al. (2009). Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc. Natl. Acad. Sci. U.S.A.* 106, 3698–3703. doi:10.1073/pnas.0812874106

Winnike, J. H., Busby, M. G., Watkins, P. B., and O'connell, T. M. (2009). Effects of a prolonged standardized diet on normalizing the human metabolome. *Am. J. Clin. Nutr.* 90, 1496–1501. doi:10.3945/ajcn.2009.28234

Wishart, D. S. (2008). Quantitative metabolomics using NMR. *Trends Analyt. Chem.* 27, 228–237. doi:10.1016/j.trac.2007.12.001

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., et al. (2013). HMDB 3.0 – the human metabolome database in 2013. *Nucleic Acids Res.* 41, D801–D807. doi:10.1093/nar/gks1065

Wishart, D. S., Lewis, M. J., Morrissey, J. A., Flegel, M. D., Jeroncic, K., Xiong, Y., et al. (2008). The human cerebrospinal fluid metabolome. *J. Chromatogr. B* 871, 164–173. doi:10.1016/j.jchromb.2008.05.001

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometr. Intell. Lab. Syst.* 2, 37–52. doi:10.1016/0169-7439(87)80084-9

Wong, J. W. H., Durante, C., and Cartwright, H. M. (2005). Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal. Chem.* 77, 5655–5661. doi:10.1021/ac050619p

Xi, Y., and Rocke, D. (2008). Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics* 9:324. doi:10.1186/1471-2105-9-324

Xia, J., Bjorndahl, T., Tang, P., and Wishart, D. (2008). MetaboMiner – semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics* 9:507. doi:10.1186/1471-2105-9-507

Xia, J., Broadhurst, D., Wilson, M., and Wishart, D. (2013). Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics* 9, 280–299. doi:10.1007/s11306-012-0482-9

Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., and Wishart, D. S. (2012). MetaboAnalyst 2.0 – a comprehensive server for metabolomic data analysis. *Nucleic Acids Res.* 40, W127–W133. doi:10.1093/nar/gks374

Xia, J., and Wishart, D. S. (2010a). MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics* 26, 2342–2344. doi:10.1093/bioinformatics/btq418

Xia, J., and Wishart, D. S. (2010b). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* 38, W71–W77. doi:10.1093/nar/gkq329

Xiao, C., Hao, F., Qin, X., Wang, Y., and Tang, H. (2009). An optimized buffer system for NMR-based urinary metabolomics with effective pH control, chemical shift consistency and dilution minimization. *Analyst* 134, 916–925. doi:10.1039/b818802e

Xie, Y., Pan, W., and Khodursky, A. B. (2005). A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* 21, 4280–4288. doi:10.1093/bioinformatics/bti685

Yang, C., He, Z., and Yu, W. (2009). Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics* 10:4. doi:10.1186/1471-2105-10-4

Yin, P., Peter, A., Franken, H., Zhao, X., Neukamm, S. S., Rosenbaum, L., et al. (2013). Preanalytical aspects and sample quality assessment in metabolomics studies of human blood. *Clin. Chem.* 59, 833–845. doi:10.1373/clinchem.2012.199257

Zhang, A., Sun, H., Wang, P., Han, Y., and Wang, X. (2012). Modern analytical techniques in metabolomics analysis. *Analyst* 137, 293–300. doi:10.1039/c1an15605e

Zhang, G., He, P., Tan, H., Budhu, A., Gaedcke, J., Ghadimi, B. M., et al. (2013). Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. *Clin. Cancer Res.* 19, 4983–4993. doi:10.1158/1078-0432.CCR-13-0209

Zhang, Z.-M., Chen, S., and Liang, Y.-Z. (2010). Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* 135, 1138–1146. doi:10.1039/b922045c

Zheng, C., Zhang, S., Ragg, S., Raftery, D., and Vitek, O. (2011). Identification and quantification of metabolites in 1H NMR spectra by Bayesian model selection. *Bioinformatics* 27, 1637–1644. doi:10.1093/bioinformatics/btr118

Zhou, B., Xiao, J. F., Tuli, L., and Ressom, H. W. (2012). LC-MS-based metabolomics. *Mol. Biosyst.* 8, 470–481. doi:10.1039/c1mb05350g

Zhu, W., and Zhang, H. (2009). Rejoinder: why do we test multiple traits in genetic association studies? *J. Korean Stat. Soc.* 38, 25–27. doi:10.1016/j.jkss.2008.10.007

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 December 2014; accepted: 18 February 2015; published online: 05 March 2015.

Citation: Alonso A, Marsal S and Julià A (2015) Analytical methods in untargeted metabolomics: state of the art in 2015. Front. Bioeng. Biotechnol. ***3****:23. doi: 10.3389/fbioe.2015.00023*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.

D.2 Metabolomics in rheumatic diseases

International Journal of Clinical Rheumatology

Review

Metabolomics in rheumatic diseases

Antonio Julià[1], Arnald Alonso[1] & Sara Marsal*[,1]

[1]Rheumatology Research Group, Vall d'Hebron Research Institute, Parc Científic de Barcelona, Torre I, 5ª Planta, c/ Baldiri i Reixac, No 4, Barcelona, 08028, Spain

*Author for correspondence: Tel.: +34 934 029 082; sara.marsal@vhir.org

Metabolites are low-weight molecules that are present in multiple biochemical processes either as intermediates or as end products of the metabolism. Consequently, the types and quantities of metabolites in a cell, tissue or organ can be informative of an underlying pathological event. Metabolomics, the global analysis of the complete metabolite profile, is a fast-developing biomedical research area. In the present article, we introduce the main methodological aspects of metabolomics and we review the most recent contributions of this approach in the study of the following rheumatic diseases: rheumatoid arthritis, systemic lupus erythematosus, ankylosing spondylitis, psoriatic arthritis, osteoarthritis and gouty arthritis.

Keywords: bioinformatics • genomics • mass spectrometry • metabolites • metabolomics • nuclear magnetic resonance • rheumatic diseases • rheumatology

The metabolome is the most dynamic level of the organism

At the molecular level, the human body is an extremely dynamic system, with thousands of molecular reactions taking place at each instant, inside millions of cells. These biochemical reactions are responsible for maintaining the cell activity, preserving cell structure and maintaining cell-to-cell communication. For example, glucose breakdown to generate the main energy transfer molecule, ATP, is performed through a series of multiple metabolic intermediates. Each metabolite, in turn, has singular physical and chemical properties that can be used to measure its concentration at a specific time point in a certain tissue or cell type. From a biomedical perspective, the characterization of the metabolomic profile of a sample obtained from a patient can be a powerful approach to identify the physiological processes that are altered in disease. This knowledge can be key in the development of new and more effective therapeutic approaches. In addition, disease-associated metabolites can be useful biomarkers with clinically relevant applications like early diagnosis or treatment personalization.

Inflammation is a complex biological process, in which vascular, immune and other tissue-specific cell types are activated to eliminate an offending agent, either an infecting microorganism or a tissue injury. Consequently, the tissue concentrations of multiple metabolites are modified from their normal homeostatic levels. An important subset of the most severe types of rheumatic diseases is characterized by the presence of chronic inflammation, leading to tissue destruction, pain, disability and the reduction of life expectancy. Identifying the metabolomic profile associated with each clinical entity would therefore be of major importance for the development of more individualized therapeutic approaches. The recent technological and methodological advances are now allowing the fast and accurate assessment of the metabolome in many different normal and pathological conditions. In the present article we will describe these technological advances and we will review the most significant results in the metabolomics study of rheumatic diseases. Supplementary Table 1 (see online at http://www.futuremedicine.com/doi/full/10.2217/IJR.14.25) summarizes the

Future Medicine part of fsg

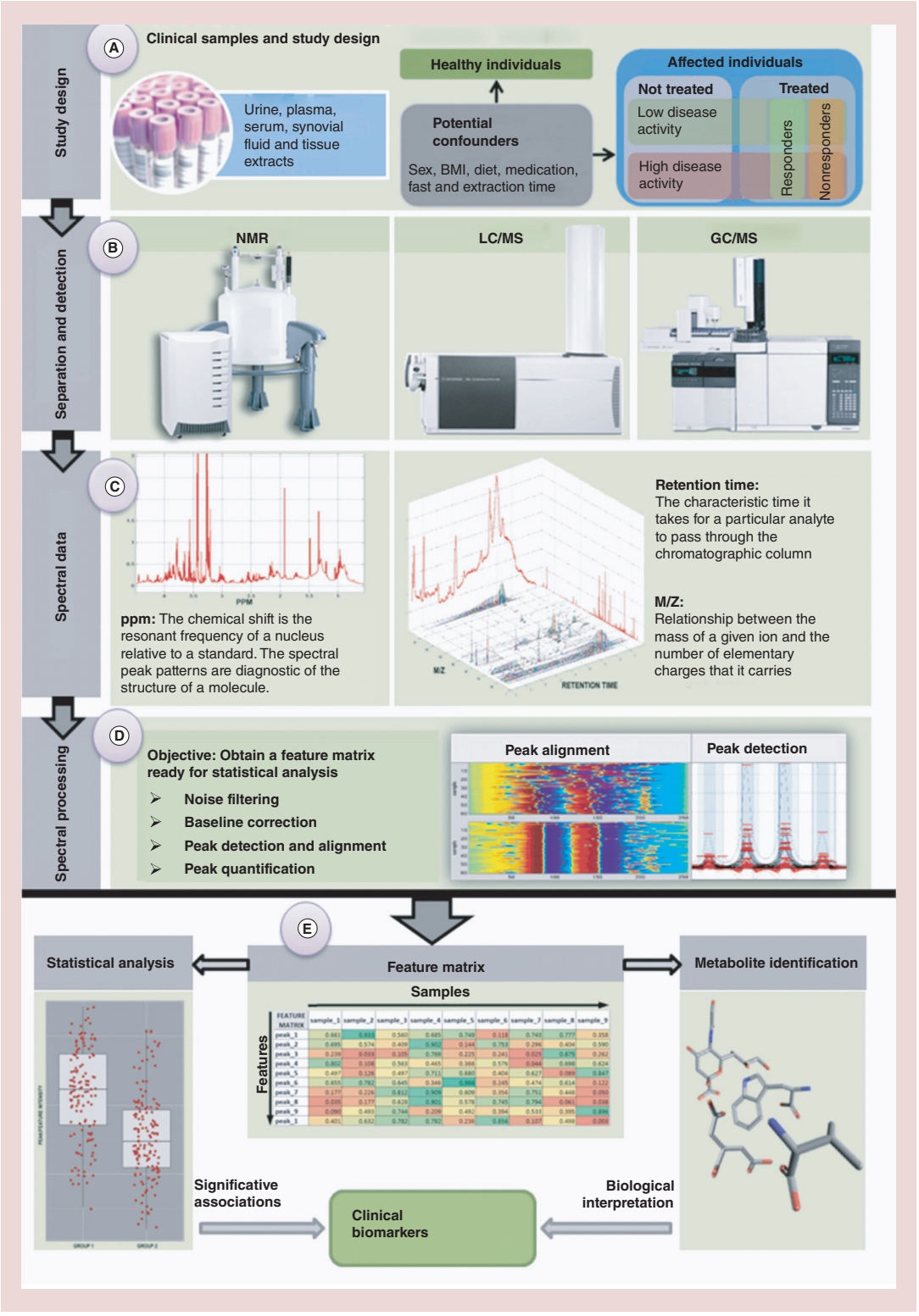

Study design
A
Clinical samples and study design
Urine, plasma, serum, synovial fluid and tissue extracts
Healthy individuals
Potential confounders
Sex, BMI, diet, medication, fast and extraction time
Affected individuals
Not treated
Treated
Low disease activity
High disease activity
Responders
Nonresponders
Separation and detection
B
NMR
LC/MS
GC/MS
Spectral data
C
PPM
ppm: The chemical shift is the resonant frequency of a nucleus relative to a standard. The spectral peak patterns are diagnostic of the structure of a molecule.
M/Z
RETENTION TIME
Retention time: The characteristic time it takes for a particular analyte to pass through the chromatographic column
M/Z: Relationship between the mass of a given ion and the number of elementary charges that it carries
Spectral processing
D
Objective: Obtain a feature matrix ready for statistical analysis
Noise filtering
Baseline correction
Peak detection and alignment
Peak quantification
Peak alignment
Peak detection
E
Statistical analysis
Feature matrix
Samples
Features
Metabolite identification
Significative associations
Clinical biomarkers
Biological interpretation

Figure 1. Metabolomics study workflow (see facing page). This figure shows the different steps associated with a metabolomics study. **(A)** According to the objectives of the study, the adequate cohorts and the type of biological samples that will be screened must be selected. **(B)** Shows the analytical instruments that are mostly used to acquire the spectral data from the biological samples. **(C)** Depending on the technological platform, the resulting spectral data are referred either in chemical shift (ppm) or in m/z and retention time. **(D)** Once the spectral data have been acquired, different processing pipelines must be applied in order to remove noise and bias and to accurately quantify each spectral peak. **(E)** This results in a feature matrix containing the quantification measures for each peak and each sample. This matrix is finally used to perform both the statistical analysis to identify significant associations and metabolite identification that will link each feature (i.e., peak) with the corresponding metabolite that will allow the biological interpretation of the identified associations.
GC/MS: Gas chromotography mass spectometry; LC/MS: Liquid chromotography mass spectometry; NMR: Nuclear magnetic resonance; ppm: Parts per million.

studies included in this review as well as the list of key metabolites identified in each disease.

Study design in metabolomic studies

There are two major study design approaches in metabolomics: targeted and untargeted studies [1,2]. In targeted studies, the researcher has a specific hypothesis to test that is based on previous knowledge of a particular biological pathway or metabolite family. In this type of studies, only a reduced set of metabolites is detected and quantified. Targeted studies are characterized for being very demanding in terms of sample preparation and analytical setup but, in exchange, metabolite measurements are sensitive and highly accurate. Consequently, these types of metabolomic studies tend to require smaller sample sizes and less bioinformatic processing steps than untargeted studies.

In untargeted studies, the goal is the measurement of the largest possible number of metabolites per sample in order to obtain a global metabolomic profile. These types of metabolomic studies are normally part of a top–down strategy, where the results obtained at the global level are used to generate new hypotheses that are subsequently validated using a targeted approach. In untargeted metabolomic studies, the large amount and complexity of the data generated require the development of highly efficient bioinformatic methods that are able to extract the most relevant biomedical information.

In metabolomics, like in many other biomedical research areas, different study designs can be considered: cross-sectional, cohort or case–control studies. Given its simplicity, one of the most commonly used study approaches in metabolomics has been the case–control design. In this design, samples from affected individuals and nonaffected individuals are drawn at random from the population at risk. With this approach, different clinically relevant comparisons can be performed to identify useful biomarkers like disease diagnostic biomarkers, disease activity biomarkers or biomarkers of drug response (Figure 1A). When using this approach in untargeted metabolomic studies of human samples, large sample sizes are generally required in order to efficiently control for potentially confounding variables like clinical and epidemiological variables (i.e., age, gender, diet or smoking status) as well as technological artifacts.

In metabolomics, the variability introduced by the observer (i.e., the biomedical researcher), can have a dramatic impact on the quality of the results. Sample manipulation, for example, is a critical step since variation at the collection, processing or sample storage steps can introduce significant biases in the resulting data. This high variability is largely due to the speed of degradation or modification of multiple metabolites in the sample. In order to minimize the negative impact of sample manipulation, the collection of human biological samples must be carefully planned and the technical variables recorded. For example, when collecting biofluids like plasma or urine the researcher must attempt to standardize influential aspects like the diet, hour of the day at which the sample is collected or the fasting time. Previous studies have shown that variation in these aspects can introduce significant biases in downstream statistical analyses [3–5]. Even seemingly irrelevant aspects like the type of containers where the samples are collected or the addition of preservatives during sample collection, can significantly alter the quality of the metabolomic data that can be obtained [4,6–8]. Consequently, the introduction of quality control measures like the standardization of all processes, the use of internal controls as well as the adequate calibration of the equipment are of paramount importance in metabolomic analysis.

Recent technological improvements are boosting metabolomic studies

In the last years, the advances in metabolomic analysis technologies as well as on bioinformatic data analysis methods have boosted the presence of metabolomics in biomedical research [2,9]. To date, the two most widely used metabolomic analysis technologies are nuclear magnetic resonance (NMR) and mass spectrometry (MS) (Figure 1B).

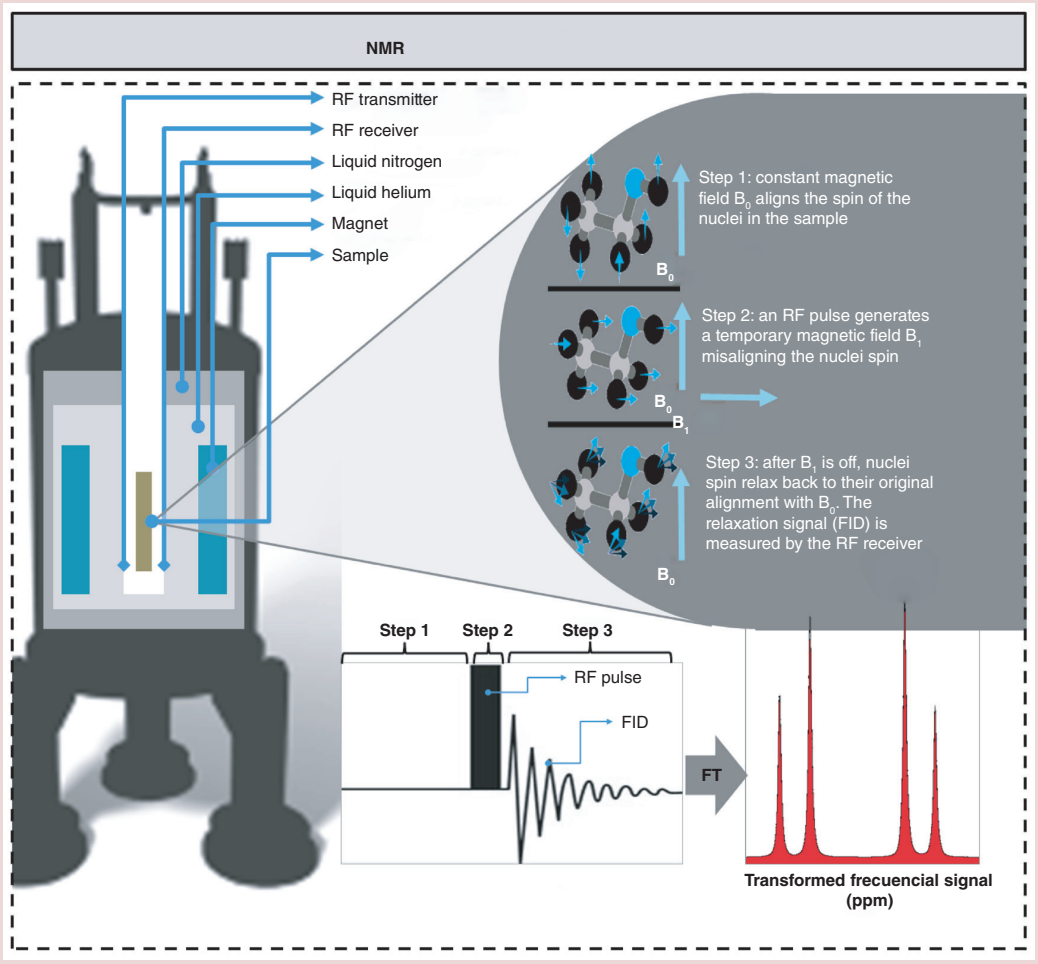


Figure 2. Analytical techniques: nuclear magnetic resonance. The NMR spectral acquisition is based on the behavior of the molecule spin under magnetic field variations. First, a constant magnetic field is applied to the sample, aligning the spins of all their molecules (i.e., step 1). The next step consists of applying a RF pulse to generate an interfering magnetic field which temporary misalign the molecules' spins (i.e., step 2). Once the interfering magnetic field disappears (i.e., step 3), the molecules' spins relax back to their original alignment. This spin relaxation results in a signal, FID, which can be measured and, after applying on the FT, is transformed on a peak spectrum where each peak is characterized by its amplitude (vertical axis) and its chemical shift (horizontal axis). The latter is usually measured in ppm and refers to the difference between the resonance frequency and that of a reference substance divided by the frequency of the spectrometer.
FID: Free induction decay; NMR: Nuclear magnetic resonance; ppm: Parts per million; RF: Radio frequency; FT: Fourier transform.

NMR is a spectroscopic analysis technique [10,11] based on the physical properties of energy absorption and re-emission of the atom nuclei due to variations in the applied magnetic field (Figure 2). Measuring the energy emitted by the atom nuclei that build up a specific molecule (i.e., free induction decay) not only allows the quantification of the concentration of the molecule itself. NMR is a fast and highly reproducible metabolomic analysis technique and has the advantage with respect to MS that it does not destroy the biological sample at study. Given its relatively low cost per sample analysis, NMR is generally the technology of choice when performing large-scale explorative studies of the metabolome.

To date, NMR-based studies have been used to identify and quantify metabolites in different types of human samples such as urine [12], serum [13] or cerebrospinal fluid [14]. Among the different NMR analytical

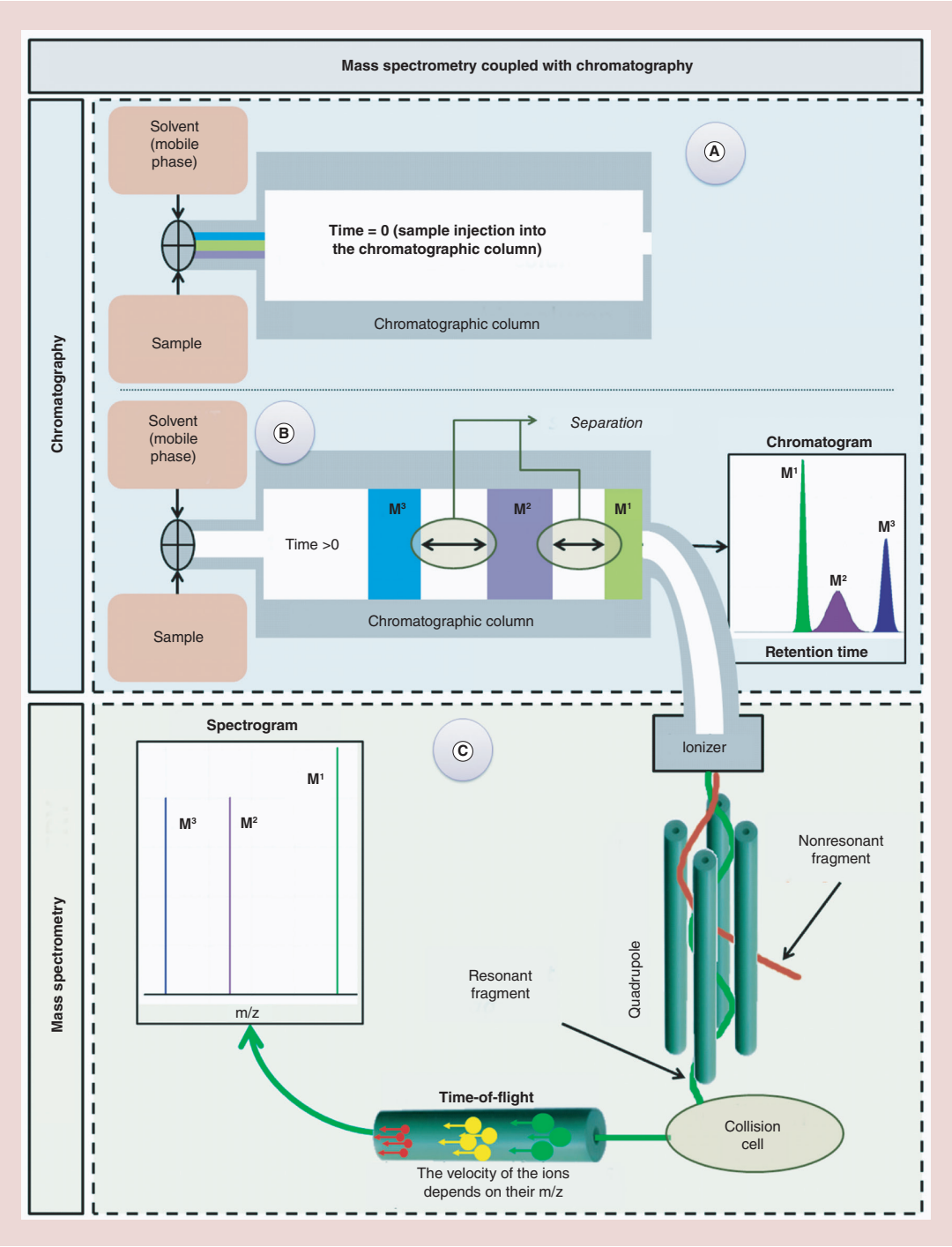


Figure 3. Analytical techniques: mass spectrometry coupled to chromatography. This figure schematizes the mass spectometry-based spectral acquisition workflow. **(A)** The sample is injected into the chromatographic column using a solvent as mobile phase. **(B)** As the sample flows through the column, the different metabolites (M^1, M^2 and M^3) are separated due to their differential retention on the stationary phase inside the column. **(C)** Once the sample has traversed the entire column, it is introduced in the mass spectrometer, which obtains the m/z of the molecules comprising the sample.
m/z: Mass-to-charge ratio.

approaches available, 1D proton NMR (^{1}H-NMR) is the most commonly used technique. The main reasons for using ^{1}H-NMR are the natural high abundance of hydrogen nuclei in metabolites, the increased sensitivity of ^{1}H-NMR compared with other NMR approaches (i.e., ^{13}C-NMR), as well as the time-efficient acquisition of spectra (see an example spectrum on Figure 1C). Together, these advantages make this technique suitable for untargeted metabolomic studies searching for biomarkers of diagnosis or prognosis, involving large sample sizes of patients and controls.

Very recently, technological advances like cryogenically cooled probes, micro-probes and increased magnetic field strength [11] are allowing a significant gain in sensitivity of NMR-based analyses. Nonetheless, NMR is unable to detect and quantify many low concentration metabolites. In order to detect and quantify this type of metabolites, MS-based technologies must be used instead.

MS is an analytical technique that generates spectral data in form of a mass-to-charge ratio (m/z) versus the relative intensity of the compounds that are generated after the ionization of the biological sample (Figure 3). The resulting ionized compounds are then measured by the MS spectrometer, generating peak signals at specific positions of the spectrum which, altogether, define the fingerprint of the original molecule. Nowadays, MS spectrometry can be performed on a broad range of instruments and techniques which use different ionization and mass selection methods [15,16]. When analyzing the metabolome, MS analysis is generally preceded by a chromatographic separation step. This step is required to reduce the high complexity of the biological sample which would otherwise be intractable. Liquid and gas chromatography columns (LC and GC, respectively) are the most commonly used chromatographic separation techniques. In both cases, metabolite separation is based on the different amount of time required by each metabolite to pass through the chromatographic column. This time, called retention time, depends on the metabolite interaction with the adsorbent material inside the column.

Compared to NMR, MS-based metabolomic analyses have a much higher sensitivity, and therefore they allow the detection and quantification metabolites that are in low concentrations in the biological samples. Compared with NMR, however, MS analyses require additional sample preparation steps [17] as well as a chromatography separation phase. For this reason, researchers performing large-scale and cost-effective exploratory studies might opt to initially use NMR techniques.

Among MS techniques, LC-MS is frequently used due to its high reproducibility and wide range of covered metabolites [18,19]. Metabolite sensitivity of LC-MS highly depends on the ionization method used (i.e., electrospray ionization is the most commonly used [20]) and also on the subsequently selected ionization mode (i.e., either positive or negative [21]). Other MS-based techniques such as GC-MS can be more reliable depending on the chemical nature of the studied metabolites (i.e., nonionizable compounds such as retinol). This technique requires the metabolites to be volatile or suited for chemical derivatization and subsequent volatilization [22,23]. Finally, the recent technological advances in LC-MS, like the use of ultra-performance liquid chromatography [20], are clearly boosting the capabilities of MS metabolomic analysis by significantly increasing specificity, sensitivity and acquisition time.

The need for bioinformatics in high-throughput metabolomic studies

The recent technological advances in metabolomic analysis have increased the quantity but also the complexity of metabolomic data that can be extracted from one single biological sample. At the same time, the development of highly specialized sample collections for 'omics' studies like biobanks are allowing the analysis of large volumes of samples. Consequently, there is a clear need for bioinformatic methods that can process this large amount of complex data and extract meaningful information in a fast and reliable way (Figure 1D).

The raw spectral data generated by NMR (Figure 1C) are known to be subject to multiple experimental and technical biases. Macro-molecule signals and other inter-sample variant factors such as used solvent, pH, ion strength and sample dilution, can produce variations on spectral peak positions and areas as well as baseline spectral artifacts [24] requiring the application of bioinformatic methods that can correct these artifacts. The two main correction methods in NMR data processing are baseline correction and peak alignment methods [25]. After these corrections are applied, two alternative bioinformatic methods can be applied to the NMR data before performing the desired statistical analysis: spectral binning and peak-based analysis. The first method is based on the automatic partitioning of the spectra in equally spaced bins. The areas calculated from each of these bins are subsequently used in the statistical analysis, in order to identify the spectral regions that are associated with the trait of interest. Limitations of this approach are that it does not account for peak positions and also that it does not exploit the presence of correlation patterns at the peak level that can be useful to identify the metabolite. In order to overcome these limitations, peak-based

methods have been developed. Making use of the correlation patterns between peaks and the information stored in the metabolite spectral databases [26], these methods can accurately identify and quantify many different metabolites present in the biological sample. In the recent years, several open-source bioinformatic methods have been developed that can partially [27,28] or completely [29] overcome all the challenges of NMR data processing.

With regards to MS-based spectral data analysis, the processing pipeline is very similar to NMR-based workflow [30]. In MS analysis, several accurate and robust bioinformatics tools have been also developed to perform MS spectra processing [31] as well as end point metabolite identification [32,33]. These tools are often dependent on the type of chromatographic column used for separation and suited for two-dimensional (i.e., retention time and m/z) spectral analysis (Figure 1C).

In addition to spectral processing and metabolite identification, high throughput metabolomic studies also require the application of complex statistical methods in order to extract relevant information (i.e., sample group clustering, biomarker identification). These multivariate analysis methods [34], also known as chemometrics methods, allow to perform tasks such as data overview, (i.e., principal component analysis), model building (i.e., partial least squares and orthogonal principal component analysis) and biomarker identification. Given the complexity associated with the metabolomics data processing and analysis, any metabolomics study will require the participation of highly trained specialists.

Metabolomic studies in rheumatoid arthritis

Since 1990, a relatively large number of studies have investigated the metabolomic changes that occur in rheumatoid arthritis (RA) pathology. Metabolomic studies in RA have evolved from using small sample sizes to investigate gross metabolic changes in changes on synovial fluid (SF) to using large sample sizes used to screen less invasive biofluids such as serum or urine, in order to disentangle the metabolome dynamics associated with disease activity and treatment response.

SF metabolome in RA

SF is a low abundance biological fluid produced from the synovial membrane filtrate of plasma and which contains high levels of hyaluronic acid. Its main purpose is the lubrication of the joint and the nutrition of the neighboring cartilage tissue. In RA, the SF is enriched with inflammatory cells, proteins and metabolites from the inflamed tissue. Therefore, the SF is a direct surrogate for the main biological processes that are taking place in the inflamed joint in RA.

The first metabolomic studies in RA were performed on SF and plasma samples from RA patients using simple NMR approaches [35–38]. The studies by Naughton *et al.* [37,38] confirmed the hypoxic nature of the inflamed synovial joint in RA, with the production of high levels of lactate compared with SF from controls. Inflammation in RA also was shown to lead to a significant consumption of lipids (i.e., reduction in low-density lipoproteins, very-low-density lipoproteins and chylomicrons), which consequently increase the concentration of ketone bodies (i.e., acetone, 3-D-hydroxybutyrate).

The anaerobic properties of RA SF were soon exploited in the search for useful disease biomarkers. Meshitsuka *et al.* [36] proposed the lactate/alanine ratio as a biomarker of RA early diagnosis since it showed increased levels compared with osteoarthritis (OA), the most common non-autoimmune arthritis. More recently, Fuchs *et al.* [39] analyzed SF and matched plasma samples of patients undergoing anti-TNF therapy. Using matrix-assisted laser desorption and ionization time-of-flight MS technology, they measured the phospholipid profile before and during anti-TNF treatment, in order to evaluate the use of plasmatic metabolites as useful markers of disease activity in the joint. They found that the ratio between phosphatidylcholine and its derivate and powerful immune chemoattractant lysophosphatidylcholine (i.e., PC/LPC ratio) that was detected in SF samples from RA patients was still detectable in their corresponding plasma samples. This aspect is of importance since it demonstrates that metabolic variations in the target tissue (i.e., SF in RA), can also be detected in plasma, which is a much less invasive type of sample. As expected, PC/LPC ratio increased with anti-TNF treatment due to the reduction of joint inflammation and the subsequent reduced production of lysophosphatidylcholine.

More recently, Giera *et al.* [40] used a more advanced technological approach to further characterize the lipid profile of RA SF. Using an untargeted LC-MS/MS system, they characterized 70 different lipid and lipid-derived metabolites. Among these metabolites they identified high levels of 5S,12S-diHETE, an isomer of LTB_4 leukotriene. 5S,12S-diHETE is produced by activated neutrophils and platelets, two cell types that have been closely related to RA pathophysiology [41,42]. The same group, using high resolution MS techniques, performed also a lipid profile analysis of RA and OA synovial samples [43]. While the unsupervised analysis of the lipid data was not able to distinguish between disease diagnostic, they were able to identify a new clustering pattern of the SF samples. A thorough analysis of this lipid profile showed that the observed

sample clustering was caused by the differential abundance of esterified oxylipids. Additional studies are needed to identify the biological origin of this family of lipids and its association with rheumatic diseases.

Serum & plasma metabolomes

In the clinical setting, blood is a convenient and useful tissue for the study of rheumatologic diseases since the collection of samples is clearly much less invasive than the sampling of the affected target tissue. In metabolomics studies of blood, the non-cellular component is separated and analyzed. When extracting the fluid component of blood, researchers can opt to use anticoagulant agents – in which case a plasma sample is obtained – or let blood clot and obtain a serum sample. The clotting of blood strongly influences the level of certain compounds like eicosanoids [13] or oxylipins. Consequently, researchers must aware of the potential impact of this non-physiologic clotting process in the results of their metabolomic studies in blood.

One of the first metabolomic approaches to the characterization of the RA serum metabolome was performed by Weljie *et al.* [44], using the K/BxN arthritic mouse model. The K/BxN model is a well-described model of inflammatory arthritis that shares many similarities with human RA. Using ^{1}H-NMR, serum samples from arthritic and control mice were compared, to identify a characteristic metabolite pattern. Uracil and TMAO were significantly increased in the serum of arthritic mice while xanthine, glycine, glycerol, hypoxanthine were significantly reduced compared to nonarthritic mice. Consequently, this study identified, for the first time, metabolites from the nucleic acid and oxidative stress pathways as potential biomarkers for RA pathology.

Using a cohort of patients and controls Lauridsen *et al.* [45] confirmed that the inflammatory state of RA is reflected in the ^{1}H-NMR spectra of human plasma samples. Similar to the discoveries in RA synovial fluid, lactate concentrations were found to be higher in patients compared with controls, probably as a consequence of the increase of the anaerobic metabolism occurring in the inflamed joints. Another likely consequence of the inflammatory activity of the synovial membrane were the high levels of acetylated glycoproteins detected in RA plasma. Finally, and in consistence with previous studies [46,47], elevated concentrations of cholesterol and low levels of high-density lipoprotein were also found to be associated with RA. High-density lipoproteins are mainly responsible for cholesterol removal from the bloodstream; together, this lipid bioprofile could explain the increased risk for coronary artery disease observed in RA patients.

Searching for a diagnostic metabolomic profile, Madsen *et al.* [48] analyzed the plasma from RA patients and compared it to controls and patients with psoriatic arthritis (PsA). In this case, a specific set of metabolites were evaluated with GC-MS and LC-MS. They observed a significant decrease of histidine levels in RA patients compared with controls and PsA patients. Low levels of this amino acid have been one of the earliest characteristic plasmatic features observed in RA [49], although its origin is still not clear. Threonic acid, a metabolite of vitamin C, was also highly expressed in RA plasma compared with controls and could be a consequence of the high oxidative stress present in the disease. Contrary to the previous study in human plasma, however, they found significantly lower levels of cholesterol in RA patients. Other studies also support the lower concentrations of cholesterol in RA compared with controls [50], so there is a clear need for additional studies to identify the potential influence of confounding variables and define the precise association of this metabolite with RA. Finally, similar to the K/BxN model screening analysis, they also found high levels of metabolites implicating the nucleotide biosynthesis (i.e., pseudouridine and guanosine) with RA. Discrepancies like the increase in hypoxanthine levels compared with the reduction in the RA mouse model remind us, however, the limitations of using animal models in the characterization of human diseases.

Very recently, Jiang *et al.* [51] used GC-MS and LC-MS to perform a metabolic analysis of the serum profile of the most prevalent forms of arthritis including RA, OA, ankylosing spondilytis (AS) and gouty arthritis (GA). Using multivariable analysis techniques on the set of measured metabolites they succeeded in discriminating all arthritis patients from healthy controls. The common arthritis profile included an increase of lactate, dihydroxyfumaric acid, glyceraldehyde, aspartic acid and homoserine, as well as a reduction in 4,8 dimethylnonanoyl carnitine. Together, the levels of these metabolites in plasma could distinguish an arthritis patient from a control individual with an 81% sensitivity and 88% specificity. In this study the differences between RA and OA (i.e., female patients) and between AS and GA (i.e., male patients) were also explored. The former comparison was based on a panel of the 13 top-ranked differential metabolites (e.g,. tryptophan, sarcosine, alanine) and yielded a classification model with an 86% sensitivity and 85% specificity. Similarly, a panel of the 16 top-ranked differential metabolites between AS and GA (e.g., creatine, cysteine, uric acid and valine) were selected to build a prediction model that reached a 79% sensitivity and 85% specificity.

The early diagnosis of RA can be crucial to the improved management of the disease and the increase

in the rates of clinical remission. For this reason, the identification of biomarkers that are informative at earlier stages of the disease can be of major importance. Using ^{1}H-NMR technology Young *et al.* [52] performed a screen in the serum metabolome of RA patients at their most initial stages and compared it to patients with more advanced disease as well as controls. Similar to Lauridsen *et al.* [45], they found low levels of lipoproteins in RA patients compared with controls. Also, the presence of 3-hydroxybutyrate, a ketone body, is in line to the previous findings of an intense anaerobic metabolism in the inflamed RA joint [37]. In the early RA group, a strong correlation of the serum metabolite profile and the degree of inflammation, represented by the levels of C-reactive protein, was found. It is therefore possible that these metabolites represent a more objective and reliable measure of the extent of the disease, including periods of apparent clinical inactivity.

Another highly relevant objective of metabolomic analysis is the identification of metabolites that are associated to disease activity and treatment response. These biomarkers could provide more objective measurements of disease activity and, therefore, allow better disease management. In their ^{1}H-NMR longitudinal analysis of human plasma, Lauridsen *et al.* [45] found that the metabolite profile of RA patients with active disease approached the profile of patients in remission after starting therapy. Importantly, both the active RA and the remission RA profiles were significantly different from the normal controls' profiles along the longitudinal study. In another longitudinal study, Madsen *et al.* [53] used GC-MS and LC-MS to identify serum metabolites correlated with the RA activity in patients starting anti-TNF therapy. In this study, the correlation of serum metabolite levels with the DAS28 disease activity score was analyzed using two independent patient cohorts. Interestingly, while highly significant linear models associated with disease activity were built in each study cohort (p-values: 6.4×10^{-6} and 9.2×10^{-3}), the predictor metabolites were quite different between both studies. A detailed analysis of the metabolite profiles of all patients suggested the existence of different underlying disease mechanisms. The existence of different RA subclasses at the molecular level has been previously suggested by whole blood transcriptomic analyses [54]. The presence of heterogeneity in RA is a clear complicating factor in the metabolomic study of the disease and consequently imposes the use of large and well-characterized patient cohorts in order to identify clinically useful biomarkers.

Urine RA metabolome

From a biomarker perspective, urine is an even more interesting biofluid than blood since it is easy to obtain and clearly non-invasive. Inflammatory diseases have shown to influence the metabolomic profile in urine [55]. Kapoor *et al.* [56] used NMR to evaluate the association of urine metabolites and the clinical outcomes of RA and PsA treated with anti-TNF therapies. For this objective, urine samples were obtained and analyzed at baseline and at 12 weeks of treatment. Multivariate analysis of the NMR spectra showed that urine metabolites could segregate RA patients with a good response from patients with a bad response to anti-TNF therapy. Among the predictor metabolites, histamine showed the highest correlation with anti-TNF response. Patients showing higher levels of this powerful cytokine were more likely to respond to anti-TNF therapy. The existence of subgroups patients showing differing levels of this metabolite could therefore explain results on its precursor histidine, which has been found to display opposing levels in different studies [48,57].

Metabolomic studies in systemic lupus erythematosus

Systemic lupus erythematosus (SLE) is an autoimmune disease characterized by a variety of clinical manifestations and a wide production of auto-antibodies [58]. This marked heterogeneity makes it a challenge for clinical specialists to diagnose SLE, particularly at the first stages of the disease. Consequently, the identification of metabolomic biomarkers that can help diagnose SLE or any of its clinical subphenotypes, would be of major importance in the management of this yet incurable disease. Studies characterizing SLE metabolomic profile are very recent and have focused on the analysis of the serum and the urine metabolome.

Serum SLE metabolome

In order to identify a metabolomic profile characteristic of SLE, 2011 Ouyang *et al.* [59] performed a ^{1}H-NMR metabolomic analysis on serum samples from SLE patients and compared it to the serum profiles of controls and patients with RA. Multivariate analysis of the ^{1}H-NMR spectra showed a higher discrimination power between both rheumatic diseases and controls than between RA and SLE. This result is in accordance with previous studies demonstrating the existence a common core of metabolites in chronic inflammatory diseases. Histidine was found to have a low concentration in the serum of both RA and SLE compared with controls. Low levels of this amino acid had been also detected in the analysis of RA plasma [48]. Previous evidence in chronic kidney disease patients suggests that low histidine levels are inversely correlated with the presence of inflammation and thus could exert anti-inflammatory and antioxidant effects [60]. Other additional amino acids (alanine, tyrosine, isoleucine, valine, phenylala-

nine, lysine, histidine and glutamine) were also found to be expressed in lower levels in RA and SLE compared with controls, possibly linked to the protein turnover associated to the high inflammatory activity occurring in both diseases. Similarly, Krebs cycle metabolites citrate and pyruvate were also in low concentrations in the serum of SLE and RA patients, suggesting an association with the increase in the energy requirements in both inflammatory diseases. As found previously in RA plasma [45], SLE serum also shows reduced levels of low-density lipoproteins. The globally altered lipid profile observed in SLE patients might be an important factor in the pathogenesis of atherosclerosis in this disease [61]. Elevated levels of lactate were also found to be a powerful discriminative biomarker between in RA patients and SLE patients or controls.

In a more recent study, Xinghong *et al.* [62] used LC-MS to analyze a subset of metabolites in sera of SLE patients and controls. Multivariate analysis of the metabolite profile confirmed the clear classification of both groups. Looking for those metabolites with a stronger ability to separate SLE patients from controls, they identified an increase of proline amino acid and several lysophosphatidyl cholines as well as a decrease of phenylalanine, tryptophyan and bilirubin. Low levels of phenylalanine had been also detected in the previous study [59], suggesting a potential use as diagnostic biomarker in the early phases of SLE.

Using also MS technologies, Wu *et al.* [63] performed a case–control study to find additional serum metabolites associated with SLE. Importantly, in this study they included an additional group of cases and controls to validate the metabolites associated with SLE in the initial (discovery) cohort. In this study they identified a marked increase in lipid peroxidation products like 9-HODE and 13-HODE. Importantly, the level of these metabolites correlated with the increase of disease activity in patients, indicating a potential applicability as biomarkers. In concordance with this rise in oxidative stress, the levels of glutathione were also reduced in SLE patients. Vitamin B6, which is necessary for the production of glutathione, was found to be significantly reduced in SLE patients' sera, which might explain the reduction of this powerful antioxidant. This finding indicates vitamin supplementation could be a potential adjunctive therapy to reduce oxidative damage in SLE. While most differential metabolites were found to be reduced in SLE patients, two metabolites associated with the leukotriene production pathway, leukotriene B4 and 5-HETE, were found to be significantly overexpressed in serum.

Urine SLE metabolome

Renal involvement is the strongest predictor of morbidity and mortality in SLE and, consequently, there is major need to identify useful biomarkers associated to this severe outcome. However, while lupus nephritis has been largely studied from a proteomics perspective [64–67], metabolomic studies on lupus nephritis are still infrequent. Recently, Romick *et al.* [68] used ^{1}H-NMR to perform a metabolomic screen of urine to identify metabolites that can help discriminate proliferative, pure membranous and focal segmental glomerulosclerosis in SLE. They found that taurine and citrate, which have been previously associated to tubular renal function [69], were strong biomarkers distinguishing proliferative (classes III/IV) from pure membranous classes (class V). While class III/IV patients had low taurine levels and normal citrate levels in urine, class V patients had low citrate levels and high taurine levels. These differences lead to an almost perfect discrimination between the two lupus nephritis subtypes. The plasmatic levels of the two metabolites are known to be regulated by the kidney. Pathological differences between both SLE nephritis subclasses might explain the differential amount of metabolite finally excreted to urine.

Brain SLE metabolome

Compared to other rheumatic diseases, the nervous system is frequently affected in SLE patients leading to neuropsychiatric syndromes [70]. An altered glucose metabolism in the brain has been associated to the development of these psychiatric symptoms in SLE patients [71]. Using a well known SLE mouse model, Alexander *et al.* [71] used ^{13}C NMR and ^{1}H-NMR to evaluate the incorporation of glucose in brain extracts compared with control mice. The results clearly confirmed the altered brain metabolism in SLE. Choline was found to be highly increased in the brains of the diseased mice. Choline is a precursor for the synthesis of phospholipids and it is known to participate in inflammation by contributing to the production of arachidonic acid [72] which, in turn, leads to the increase of prostaglandins which exert multiple roles in the inflammatory response [73]. Glutamate and glutamine were also significantly increased and, together, suggest a predominant role for the glial cells (astrocytes and microglia) rather than neurons in the pathological events in the SLE brain. Also, lactate levels were found to be increased; lactate might be a product of infiltrating macrophages in the inflamed brain and it could also contribute to the alteration of the brain functionality in SLE.

Metabolomic studies in AS

AS is a chronic inflammatory disease characterized by axial skeleton ankylosis, enthesitis inflammation and, occasionally, peripheral arthritis. AS has an overall incidence between 0.5 and 14 per 100,000 people per

year and is more common in men [74]. Metabolomic studies in AS are very recent and have been all performed in blood samples with the final objective of detecting diagnostic biomarkers [51,75,76].

AS blood metabolome

Gao *et al.* [76] performed a case–control study using GC-MS an LC-MS to identify AS biomarkers in plasma samples. Supervised partial least squares discriminant analysis was able to accurately discriminate both samples groups demonstrating the potential of metabolomics as a reliable diagnostic tool in AS. In this study, proline, glucose, phosphate, phenylalanine, urea, glycerol and homocysteine were detected at higher concentrations in AS patients than in healthy controls. Instead, propanoic acid, tryptophan and several phosphatidylcholines were present at lower concentrations in AS patients compared with controls. Tryptophan reduction might respond to the indoleamine 2,3-dioxygenase enzyme activation by the high levels of interferon gamma produced by the disease. AS patients responding to anti-inflammatory treatments have shown to increase this amino acid and, consequently, it could become a useful biomarker of disease activity. Cartilage breakdown by the chronic inflammation in AS could explain the observed high levels of proline. Consequently, the high levels of urea would therefore caused by the rise in this and additional amino acids detected in AS patients.

Using LC-MS, Fischer *et al.* [75] also performed a case–control study with serum samples of AS patients and healthy controls. Like Gao *et al.* [76], multivariate analysis on the measured metabolomic profile was able to distinguish patients from controls. Importantly, the metabolomic data could also separate AS patients according to the Bath Ankylosing Spondylitis Disease Activity Index (i.e., BASDAI). Although most of the associated molecular features detected by LC-MS were not linked to known metabolites, they were able to identify 25-hydroxyvitamin D3 26,23-peroxylactone as a metabolite clearly downregulated in AS. The lower levels of this metabolite might reflect an alteration in the vitamin D3 metabolism, which has been shown to have profound effects in bone remodeling and immune cell activation [77]. Consequently, targeting this biological pathway could have protective effects in the bone destruction process associated to AS.

Finally, the arthritis screening study performed by Jiang *et al.* [51] provided additional support to the utility of serum as a useful surrogate of AS pathology. Using GC-MS and LC-MS on male subjects, they were able to identify several metabolites that could efficiently distinguish AS from GA patients. Creatinine, uric acid, arabitol, succinic acid, valine and 5-oxoproline were among the metabolites found to be significantly underexpressed in AS serum compared with GA serum.

Metabolomic studies in PsA

PsA is an inflammatory arthritis that is associated with psoriasis. PsA has specific clinical features such as arthritis of the distal interphalangeal joints to spondylitis and sacroiliitis. PsA occurs in approximately 12% of psoriasis patients [78] and it is associated with higher morbidity and mortality and also requires a markedly different therapeutic approach. Therefore, the identification of metabolites that can characterize PsA from purely cutaneous psoriasis could become a useful clinical tool. To date, however, no studies have directly compared PsA and psoriasis metabolomics profiles.

PsA blood metabolome

In their MS screen of RA, Madsen *et al.* [48] included a PsA cohort as an additional control group to evaluate the specificity of the metabolomic markers. Interestingly, the metabolomic profiles from RA patients were found to be more different between RA and PsA patients than RA and control patients. In this study, however, no direct contrast between PsA and controls was performed. Consequently, the identification of metabolites specifically associated with PsA can only be extrapolated from the differences between RA and controls and RA and PsA patients. Several amino acids including aspartic acid, glutamic acid, glutamate and serine were increased in PsA patients compared with RA and were not different between RA and controls. Therefore, it is likely that PsA is characterized by a major protein turnover and a higher increase of free amino acids in serum. However, specific case–control design studies must still be carried out to confirm the specificity of this metabolomic profile in PsA.

PsA urine metabolome

In their NMR longitudinal study of the urine metabolite profile associated to the response to anti-TNF treatment, Kapoor *et al.* [56] also included a cohort of PsA patients. Like for RA patients, the urine metabolome in PsA correlated with the changes in disease activity induced by the biologic treatment. The metabolite levels influenced by anti-TNF treatment were found to be similar between both diseases. In particular, high levels of glutamine, phenylacetic acid and histamine in the baseline urine samples were found to be predictors of the good response to anti-TNF treatment.

Metabolomic studies in OA

OA is the most common type of arthritis and is a major cause of pain and disability in the elderly [79]. The clin-

ical and radiologic findings that form the basis of the diagnosis of OA are poorly sensitive for monitoring the progression of the disease. Consequently, the identification of metabolites that can better reflect quantitative and dynamic changes of the joint tissue turnover would be of major utility in daily clinical practice.

OA urine metabolome

A first approach to the characterization of metabolites associated to OA pathology was performed by Lamers *et al.* analyzing the urine profiles of animal models with ^{1}H-NMR [80]. The Hartley outbred strain of guinea pigs has shown to develop progressive knee OA and, consequently, are a useful animal model to screen for potential biomarkers in OA. Principal component analysis, a multivariate approach commonly used in genomic studies, showed a clear separation between the urine profile of the OA model and the urine profile of the control strain. The major changes associated with OA were found in lactate, malate, hypoxanthine and alanine levels, which support the hypothesis that in OA there is an increase in energy utilization and an altered metabolism of purines. Having proven the existence of a urine metabolite pattern correlated with the presence of the disease, the same group performed a study using a cohort of OA patients and matched controls. Using the same multivariate analysis approach as in the animal model they were also able to discriminate between patients and controls. The prediction model built from the metabolite concentrations showed a strong correlation with the Kellgren–Lawrence radiographic scores for the evaluation of OA severity. Additionally, the metabolite profile that characterized the human OA patients shared many features with the profile obtained in the guinea pig model, confirming the usefulness of this model to pursue clinically relevant biomarkers. Like in the animal model, several of the differential NMR signals could not be identified. This aspect is characteristic of the exploratory nature of ^{1}H-NMR; in these cases, additional studies using more sensitive technologies like correlation spectroscopy NMR, LC-MS or GS-MS are required to characterize the associated metabolites. The increased metabolites in OA that could be identified with certainty were hydroxybutyrate, pyruvate, creatine/creatinine and glycerol. These results suggest the increased use of fat as an energy source in OA. Histidine and methylhistidine were found to be in significant lower concentration in the urine OA patients compared with controls. Low levels of histidine in OA could be associated to an over expression of histidine decarboxylase in OA chondrocytes and the subsequent increase in the production of histamine observed in the OA joint [81]. There is evidence supporting that histamine promotes the formation of the chondrocyte clusters characteristic of the osteoarthritic cartilage [82].

OA blood metabolome

Using ^{1}H-NMR Zhai *et al.* [83] analyzed the metabolomics profile in the serum of OA patients and controls. Importantly, in this study an independent replication cohort of patients and controls was also recruited to validate the metabolite associations identified in the discovery phase. Also, in this study they also used the ratios between metabolites since it has been shown to provide an improved quantification of some of the metabolic reactions present in the tissue of interest. After correcting for the number of statistical tests performed, they found the valine/histidine and the xleucine(isoleucine and leucine)/histidine ratios to be significantly associated with the presence of knee OA. These ratios showed also to be predictive of the OA severity, showing a correlation with the Kellgren–Lawrence OA grade. Interestingly, valine, isoleucine and leucine belong to the branched-chain family of amino acids (BCAA). BCAA are characterized for being essential amino acids (i.e., they cannot be synthesized by the body), having a similar molecular structure and being important constituents of the skeletal muscle. Elevated levels of BCAA have been also found both associated to obesity [84] and aging [85]. Importantly, BMI and age were discarded as confounders from the observed association between the two ratios and knee OA. Another potential explanation for the increased levels of these amino acids could be the collagen breakdown that is associated with this disease. A recent study comparing the metabolomic profile of media conditioned by cultured synovial fibroblasts from OA patients and controls also found additional evidence implicating the BCAA metabolism with OA etiology [86]. Consequently, the BCAA/histidine ratio could become a valuable biomarker in the management of OA.

Metabolomic studies in GA

Approximately 1% of the adult men in western countries have gout. GA is characterized by recurrent attacks of acute monoarticular or oligoarticular inflammation caused by the formation of urate crystals in the joint [87]. Misdiagnosis in the early stages of the disease can influence the outcome of the disease. Consequently, there is a need to identify biomarkers that can help to reliably diagnose the disease.

GA metabolome

Liu *et al.* [88] used high performance liquid chromatography to analyze both the serum and urine profiles of GA patients and matched controls. The multivariate

analysis of the obtained profiles was able to discriminate between the case and controls groups. Patients with GA showed increased serum levels of uric acid, creatinine and tryptophan. Uric acid is the end product of purine degradation and, in high concentrations, it leads to the formation of monosodium urate crystals in the joint. Although this is a well established pathological process of GA, this study shows that while it uric acid could be an informative biomarker at the earlier stages of the disease, it is not sufficiently informative

Executive summary

The metabolome is the most dynamic level of the organism

- The characterization of the metabolomic profile of patients can be a powerful approach to identify the physiological processes that are altered in disease.
- The recent technological and methodological advances are now allowing the fast and accurate assessment of the metabolome.

Study design in metabolomic studies

- There are two major study design approaches in metabolomics: targeted and untargeted studies.
- Given its simplicity, one of the most commonly used study approaches in metabolomics has been the case–control design.
- In metabolomics, the technical and biological variabilites can have a dramatic impact on the quality of the results and must be controlled.

Recent technological improvements are boosting metabolomic studies

- The two most widely used metabolomic analysis technologies are nuclear magnetic resonance (NMR) and mass spectrometry (MS).
- NMR is a fast and highly reproducible metabolomic analysis.
- MS has a higher sensitivity than NMR, but is more demanding in terms of sample preparation and technical requirements.

The need for bioinformatics in high-throughput metabolomic studies

- There is a clear need for bioinformatic methods that can process these large amounts of complex data in metabolomic studies and extract meaningful information.
- In the recent years, several open-source bioinformatic methods have been developed that can overcome several challenges of NMR and MS data processing.

Metabolomic studies in rheumatoid arthritis

- Metabolomic studies confirm the hypoxic nature of the inflamed synovial joint in rheumatoid arthritis (RA).
- Metabolite profiles in RA plasma are associated to disease activity and treatment response.
- Metabolites from the nucleic acid and oxidative stress pathways are potential biomarkers for RA pathology.

Metabolomic studies in systemic lupus erythematosus

- Low histidine levels have been identified in the serum of systemic lupus erythematosus (SLE) and RA.
- The commonly altered lipid profile in SLE and RA could explain the increased incidence of cardiovascular disease observed for both rheumatic diseases.
- Taurine and citrate levels in the SLE urine metabolome have potential utility as biomarkers for SLE nephritis subtype discrimination.

Metabolomic studies in ankylosing spondylitis

- Metabolomic studies of patients and controls identified reduced levels of tryptophan in ankylosing spondylitis, probably due to IFN-γ expression in the disease.
- Metabolite levels in serum reflect an alteration in the vitamin D3 metabolism in ankylosing spondylitis.

Metabolomic studies in psoriatic arthritis

- To date, no studies have directly compared psoriatic arthritis and psoriasis metabolomics profiles.
- The urine metabolome in psoriatic arthritis is correlated with the changes in disease activity induced by anti-TNF treatment.

Metabolomic studies in osteoarthritis

- The prediction model built from the urine metabolite concentrations correlates significantly with the Kellgren–Lawrence radiographic scores of osteoarthritis severity.
- The valine/histidine and the xleucine(isoleucine and leucine)/histidine ratios are potential biomarkers of the development of knee osteoarthritis.

Metabolomic studies in gouty arthritis

- Uric acid is not a sufficiently informative biomarker of gouty arthritis and additional markers must be identified.
- Gouty arthritis also expresses the common core of serum metabolites found in other prevalent arthritis. This common set of metabolites could be useful for the development of improved diagnostic systems.

to be used as a single diagnostic marker. Creatinine, a widely used biomarker of the renal function, could be associated to the renal affection by the deposition of urate crystals. Compared to RA, AS or SLE, where tryptophan serum levels are significantly reduced, it appears to be significantly increased in GA. The origin of this variation is not clear and, while it could be associated to GA pathology, it could also be a metabolism product resulting from the treatment received by the patients. Therapies can therefore be a major confounder in metabolomic studies if inadequately controlled.

Interestingly, uric acid and creatinine levels in urine showed an opposite variation compared with serum of GA patients. This inverse relation could be explained by the defects of tubular secretion associated with this disease. Hippuric acid, a conjugate of benzoic acid and glycine normally generated by microfloral metabolism, was also found to be reduced in GA urine. The authors, however, suggest that the increased energy consumption associated with GA associated inflammation could be responsible for this observed metabolite reduction.

In their recent screen of the serum metabolomic profiles of four different types of arthritis, Jiang [51], confirmed the high diagnostic utility of uric acid and creatine for GA. Also, additional metabolites like cystine, arabitol and alloxanoic acid were found to be in high concentrations in patients with GA. Used together in a multivariate model, they could clearly distinguish GA patients from AS patients or controls. The results of this study are a strong basis for the development of diagnostic systems based on the screening of multiple informative biomarkers.

Future perspective

Recent technological advances have boosted the capacity to mine the metabolite composition of biological samples associated with different diseases. This new layer of information strongly complements the previously established genomic, transcriptomic and proteomic technologies. In the near future, studies integrating these different layers of biological information will provide essential knowledge for the identification of the biological mechanisms that operate in each disease and will provide an accurate molecular profile of each patient. This individual profile will have a high translational potential in rheumatic disease since it could help to advance the time of diagnosis as well as help medical specialists to perform more guided therapeutic decisions.

There are still, however, several challenges that need to be overcome. The annotation of many metabolites must clearly evolve, a task that is actually being carried out by different databases [26,89,90]. Metabolome analysis technologies must improve in sensitivity and be less time consuming and costly and the associated analysis algorithms must improve their accuracy. Also, if large cohort analyses are to be performed, there is a clear need for improvement in the throughput of most metabolomics platforms. Additionally, sample and clinical collection procedures must be standardized to ensure the quality of the results and the minimization of technical and biological confounders. Although metabolomics is an emergent discipline, it is rapidly evolving and, in the next years, new findings will clearly increase our knowledge of the molecular basis of rheumatic diseases and contribute to improve the prognosis of these patients.

Financial & competing interests disclosure

This work was supported by the Spanish Ministry of Economy and Competitiveness, grants PI12/01362, PSE-010000-2006-6 and IPT-010000-2010-36, and by the AGAUR FI grant (2013/00974). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1 Dunn WB, Broadhurst DI, Atherton HJ, Goodacre R, Griffin JL. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem. Soc. Rev.* 40(1), 387–426 (2011).

2 Patti GJ, Yanes O, Siuzdak G. Innovation: metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* 13(4), 263–269 (2012).

3 Rasmussen L, Savorani F, Larsen T, Dragsted L, Astrup A, Engelsen S. Standardization of factors that influence human urine metabolomics. *Metabolomics* 7(1), 71–83 (2011).

4 Townsend MK, Clish CB, Kraft P *et al.* Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. *Clin. Chem.* 59(11), 1657–1667 (2013).

5 Winnike JH, Busby MG, Watkins PB, O'Connell TM. Effects of a prolonged standardized diet on normalizing the human metabolome. *Am. J. Clin. Nutr.* 90(6), 1496–1501 (2009).

6 Bando K, Kawahara R, Kunimatsu T *et al.* Influences of biofluid sample collection and handling procedures on GC-MS based metabolomic studies. *J. Biosci. Bioeng.* 110(4), 491–499 (2010).

7 Lauridsen M, Hansen SH, Jaroszewski JW, Cornett C. Human urine as test material in 1H NMR-based metabonomics: recommendations for sample preparation and storage. *Anal. Chem.* 79(3), 1181–1186 (2007).

8 Want EJ, Wilson ID, Gika H *et al.* Global metabolic profiling procedures for urine using UPLC-MS. *Nat. Protoc.* 5(6), 1005–1018 (2010).

9 Zhang A, Sun H, Wang P, Han Y, Wang X. Modern analytical techniques in metabolomics analysis. *Analyst* 137(2), 293–300 (2012).

10 Bothwell JH, Griffin JL. An introduction to biological nuclear magnetic resonance spectroscopy. *Biol. Rev. Camb. Philos. Soc.* 86(2), 493–510 (2011).

11 Emwas A-H, Salek R, Griffin J, Merzaban J. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics* 9(5), 1048–1072 (2013).

12 Bouatra S, Aziat F, Mandal R *et al.* The human urine metabolome. *PLoS ONE* 8(9), e73076 (2013).

13 Psychogios N, Hau DD, Peng J *et al.* The human serum metabolome. *PLoS ONE* 6(2), e16957 (2011).

14 Wishart DS, Lewis MJ, Morrissey JA *et al.* The human cerebrospinal fluid metabolome. *J. Chromatogr. B* 871(2), 164–173 (2008).

15 El-Aneed A, Cohen A, Banoub J. Mass spectrometry, review of the Basics: electrospray, MALDI, and commonly used mass analyzers. *Appl. Spectrosc. Rev.* 44(3), 210–230 (2009).

16 Marshall AG, Hendrickson CL. High-resolution mass spectrometers. *Annu. Rev. Anal. Chem.* 1(1), 579–599 (2008).

17 Vuckovic D. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography–mass spectrometry. *Anal. Bioanal. Chem.* 403(6), 1523–1548 (2012).

18 Gika HG, Theodoridis GA, Wilson ID. Liquid chromatography and ultra-performance liquid chromatography–mass spectrometry fingerprinting of human urine: sample stability under different handling and storage conditions for metabonomics studies. *J. Chromatogr. A* 1189(1–2), 314–322 (2008).

19 Theodoridis GA, Gika HG, Want EJ, Wilson ID. Liquid chromatography–mass spectrometry based global metabolite profiling: a review. *Anal. Chim. Acta* 711(0), 7–16 (2012).

20 Denoroy L, Zimmer L, Renaud B, Parrot S. Ultra high performance liquid chromatography as a tool for the discovery and the analysis of biomarkers of diseases: a review. *J. Chromatogr. B* 927(0), 37–53 (2013).

21 Nordström A, Want E, Northen T, Lehtiö J, Siuzdak G. Multiple ionization mass spectrometry strategy used to reveal the complexity of metabolomics. *Anal. Chem.* 80(2), 421–429 (2007).

22 Koek M, Jellema R, Greef J, Tas A, Hankemeier T. Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics* 7(3), 307–328 (2011).

23 Tsugawa H, Bamba T, Shinohara M, Nishiumi S, Yoshida M, Fukusaki E. Practical non-targeted gas chromatography/mass spectrometry-based metabolomics platform for metabolic phenotype analysis. *J. Biosci Bioeng.* 112(3), 292–298 (2011).

24 Xiao C, Hao F, Qin X, Wang Y, Tang H. An optimized buffer system for NMR-based urinary metabonomics with effective pH control, chemical shift consistency and dilution minimization. *Analyst* 134(5), 916–925 (2009).

25 Zhang S, Nagana Gowda GA, Ye T, Raftery D. Advances in NMR-based biofluid analysis and metabolite profiling. *Analyst* 135(7), 1490–1498 (2010).

26 Wishart DS, Jewison T, Guo AC *et al.* HMDB 3.0 – the Human Metabolome Database in 2013. *Nucleic Acids Res.* 41(D1), D801–D807 (2013).

27 Hao J, Astle W, De Iorio M, Ebbels TMD. BATMAN – an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* 28(15), 2088–2090 (2012).

28 Ludwig C, Gunther U. MetaboLab – advanced NMR data processing and analysis for metabolomics. *BMC Bioinform.* 12(1), 366 (2011).

29 Alonso A, Rodríguez MA, Vinaixa M *et al.* Focus: a robust workflow for one-dimensional NMR spectral analysis. *Anal. Chem.* 86(2), 1160–1169 (2013).

30 Castillo S, Gopalacharyulu P, Yetukuri L, Orešič M. Algorithms and tools for the preprocessing of LC–MS metabolomics data. *Chemometr. Intel. Lab. Syst.* 108(1), 23–32 (2011).

31 Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78(3), 779–787 (2006).

32 Alonso A, Julià A, Beltran A *et al.* AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics* (2011).

33 Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* 84(1), 283–289 (2011).

34 Madsen R, Lundstedt T, Trygg J. Chemometrics in metabolomics – a review in human disease diagnosis. *Analyt. Chim. Acta* 659(1–2), 23–33 (2010).

35 Claxson A, Grootveld M, Chander C *et al.* Examination of the metabolic status of rat air pouch inflammatory exudate by high field proton NMR spectroscopy. *Biochim. Biophys. Acta (BBA)* 1454(1), 57–70 (1999).

36 Meshitsuka S, Yamazaki E, Inoue M, Hagino H, Teshima R, Yamamoto K. Nuclear magnetic resonance studies of synovial fluids from patients with rheumatoid arthritis and osteoarthritis. *Clin. Chim. Acta* 281(1–2), 163–167 (1999).

37 Naughton D, Whelan M, Smith EC, Williams R, Blake DR, Grootveld M. An investigation of the abnormal metabolic status of synovial fluid from patients with rheumatoid arthritis by high field proton nuclear magnetic resonance spectroscopy. *FEBS Lett.* 317(1–2), 135–138 (1993).

38 Naughton DP, Haywood R, Blake DR, Edmonds S, Hawkes GE, Grootveld M. A comparative evaluation of the metabolic profiles of normal and inflammatory knee-joint synovial fluids by high resolution proton NMR spectroscopy. *FEBS Lett.* 332(3), 221–225 (1993).

39 Fuchs B, Schiller J, Wagner U, Häntzschel H, Arnold K. The phosphatidylcholine/lysophosphatidylcholine ratio in human plasma is an indicator of the severity of rheumatoid arthritis:

investigations by 31P NMR and MALDI-TOF MS. *Clin. Biochem.* 38(10), 925–933 (2005).

- **Proves the utility of plasma as a useful surrogate tissue to capture the activity of rheumatoid arthritis (RA).**

40 Giera M, Ioan-Facsinay A, Toes R *et al.* Lipid and lipid mediator profiling of human synovial fluid in rheumatoid arthritis patients by means of LC-MS/MS. *Biochim Biophys Acta.* 1821(11), 1415–1424 (2012).

- **This exhaustive lipid metabolite profiling study is a powerful basis for the development of biomarkers that are representative the disease activity characteristic of RA.**

41 Borgeat P, Fruteau De Laclos B, Picard S, Drapeau J, Vallerand P, Corey EJ. Studies on the mechanism of formation of the 5S, 12S-dihydroxy-6,8,10,14(E,Z,E,Z)-icosatetraenoic acid in leukocytes. *Prostaglandins* 23(5), 713–724 (1982).

42 Marcus AJ, Broekman MJ, Safier LB *et al.* Formation of leukotrienes and other hydroxy acids during platelet-neutrophil interactions *in vitro*. *Biochem. Biophys. Res. Commun.* 109(1), 130–137 (1982).

43 Jónasdóttir HS, Nicolardi S, Jonker W *et al.* Detection and structural elucidation of esterified oxylipids in human synovial fluid by electrospray ionization-fourier transform ion-cyclotron mass spectrometry and liquid chromatography-ion trap-MS3: detection of esterified hydroxylated docosapentaenoic acid containing phospholipids. *Anal. Chem.* 85(12), 6003–6010 (2013).

44 Weljie AM, Dowlatabadi R, Miller BJ, Vogel HJ, Jirik FR. An inflammatory arthritis-associated metabolite biomarker pattern revealed by 1H NMR spectroscopy. *J. Proteome Res.* 6(9), 3456–3464 (2007).

- **Identified, for the first time, metabolites from the nucleic acid and oxidative stress pathways as potential biomarkers for RA pathology.**

45 Lauridsen MB, Bliddal H, Christensen R *et al.* 1H NMR spectroscopy-based interventional metabolic phenotyping: a cohort study of rheumatoid arthritis patients. *J. Proteome Res.* 9(9), 4545–4553 (2010).

46 Dursunoğlu D, Evrengül H, Polat B *et al.* Lp(a) lipoprotein and lipids in patients with rheumatoid arthritis: serum levels and relationship to inflammation. *Rheumatol. Int.* 25(4), 241–245 (2005).

47 Georgiadis A, Papavasiliou E, Lourida E *et al.* Atherogenic lipid profile is a feature characteristic of patients with early rheumatoid arthritis: effect of early treatment – a prospective, controlled study. *Arthr. Res. Ther.* 8(3), R82 (2006).

48 Madsen RK, Lundstedt T, Gabrielsson J *et al.* Diagnostic properties of metabolic perturbations in rheumatoid arthritis. *Arthr. Res Ther.* 13(1), R19 (2011).

49 Gerber DA. Low free serum histidine concentration in rheumatoid arthritis. A measure of disease activity. *J. Clin. Invest.* 55(6), 1164–1173 (1975).

50 Liao KP, Cai T, Gainer VS *et al.* Lipid and lipoprotein levels and trend in rheumatoid arthritis compared to the general population. *Arthr. Care Res.* 65(12), 2046–2050 (2013).

51 Jiang M, Chen T, Feng H *et al.* Serum metabolic signatures of four types of human arthritis. *J. Proteome Res.* 12(8), 3769–3779 (2013).

- **Results of this study indicate a common core of metabolic pathways associated with arthritis, which are explained by common pathogenic mechanisms such as hypoxia.**

52 Young SP, Kapoor SR, Viant MR *et al.* The impact of inflammation on metabolomic profiles in patients with arthritis. *Arthr. Rheum.* 65(8), 2015–2023 (2013).

53 Madsen R, Rantapää-Dahlqvist S, Lundstedt T, Moritz T, Trygg J. Metabolic responses to change in disease activity during tumor necrosis factor inhibition in patients with rheumatoid arthritis. *J. Proteome Res.* 11(7), 3796–3804 (2012).

54 Van Der Pouw Kraan TCTM, Wijbrandts CA, Van Baarsen LGM *et al.* Rheumatoid arthritis subtypes identified by genomic profiling of peripheral blood cells: assignment of a type I interferon signature in a subpopulation of patients. *Ann. Rheum. Dis.* 66(8), 1008–1014 (2007).

55 Stephens NS, Siffledeen J, Su X, Murdoch TB, Fedorak RN, Slupsky CM. Urinary NMR metabolomic profiles discriminate inflammatory bowel disease from healthy. *J. Crohn's Col.* 7(2), e42–e48 (2013).

56 Kapoor SR, Filer A, Fitzpatrick MA *et al.* Metabolic profiling predicts response to anti-tumor necrosis factor alpha therapy in patients with rheumatoid arthritis. *Arthr. Rheum.* 65(6), 1448–1456 (2013).

- **Demonstrates the utility of urine to develop biomarkers for personalized treatment in psoriatic arthritis and RA.**

57 Sitton NG, Dixon JS, Bird HA, Wright V. Serum biochemistry in rheumatoid arthritis, seronegative arthropathies, osteoarthritis, SLE and normal subjects. *Rheumatology* 26(2), 131–135 (1987).

58 Danchenko N, Satia JA, Anthony MS. Epidemiology of systemic lupus erythematosus: a comparison of worldwide disease burden. *Lupus* 15(5), 308–318 (2006).

59 Ouyang X, Dai Y, Wen JL, Wang LX. 1H NMR-based metabolomic study of metabolic profiling for systemic lupus erythematosus. *Lupus* 20(13), 1411–1420 (2011).

- **Commonly altered lipid profile in systemic lupus erythematosus and RA could explain the increased incidence of cardiovascular disease observed for both rheumatic diseases.**

60 Watanabe M, Suliman ME, Qureshi AR *et al.* Consequences of low plasma histidine in chronic kidney disease patients: associations with inflammation, oxidative stress, and mortality. *Am. J. Clin. Nutr.* 87(6), 1860–1866 (2008).

61 De Carvalho JF, Bonfá E, Borba EF. Systemic lupus erythematosus and "lupus dyslipoproteinemia". *Autoimmun. Rev.* 7(3), 246–250 (2008).

62 Xinghong D, Ying Y. Serum metabonomics study of systemic lupus erythematosus by rapid resolution liquid chromatography coupled with Q-TOF mass spectrometry. Presented at: *Biomedical Engineering and Biotechnology (iCBEB), 2012 International Conference*. Macau, China, 28–30 May 2012.

63 Wu T, Xie C, Han J *et al.* Metabolic disturbances associated with systemic lupus erythematosus. *PLoS ONE* 7(6), e37210 (2012).

64 Li Y, Tucci M, Narain S *et al.* Urinary biomarkers in lupus nephritis. *Autoimmun. Rev.* 5(6), 383–388 (2006).

65 Mosley K, Tam FWK, Edwards RJ, Crozier J, Pusey CD, Lightstone L. Urinary proteomic profiles distinguish between active and inactive lupus nephritis. *Rheumatology* 45(12), 1497–1504 (2006).

66 Oates JC, Varghese S, Bland AM *et al.* Prediction of urinary protein markers in lupus nephritis. *Kidney Int.* 68(6), 2588–2592 (2005).

67 Zhang X, Jin M, Wu H *et al.* Biomarkers of lupus nephritis determined by serial urine proteomics. *Kidney Int.* 74(6), 799–807 (2008).

68 Romick-Rosendale L, Brunner H, Bennett M *et al.* Identification of urinary metabolites that distinguish membranous lupus nephritis from proliferative lupus nephritis and focal segmental glomerulosclerosis. *Arthr. Res. Ther.* 13(6), R199 (2011).

69 Christians U, Schmitz V, Schoning W, Bendrick-Peart J, Klawitter J, Haschke M. Toxicodynamic therapeutic drug monitoring of immunosuppressants: promises, reality, and challenges. *Ther. Drug Monit.* 30(2), 151–158 (2008).

70 Brey RL, Holliday SL, Saklad AR *et al.* Neuropsychiatric syndromes in lupus: prevalence using standardized definitions. *Neurology* 58(8), 1214–1220 (2002).

71 Alexander JJ, Zwingmann C, Quigg R. MRL/lpr mice have alterations in brain metabolism as shown with [1H–13C] NMR spectroscopy. *Neurochem. Intern.* 47(1–2), 143–151 (2005).

72 Kaiser E, Chiba P, Zaky K. Phospholipases in biology and medicine. *Clin. Biochem.* 23(5), 349–370 (1990).

73 Korotkova M, Jakobsson P-J. Persisting eicosanoid pathways in rheumatic diseases. *Nat. Rev. Rheumatol.* 10(4), 229–241 (2014).

74 Braun J, Sieper J. Ankylosing spondylitis. *Lancet* 369(9570), 1379–1390

75 Fischer R, Trudgian DC, Wright C *et al.* Discovery of candidate serum proteomic and metabolomic biomarkers in ankylosing spondylitis. *Mol. Cell Proteom.* 11(2), M111.013904 (2012).

• **Shows the utility of the plasmatic metabolome to differentiate ankylosing spondylitis patients from controls as well ankylosing spondylitis patients according to disease activity.**

76 Gao P, Lu C, Zhang F *et al.* Integrated GC-MS and LC-MS plasma metabonomics analysis of ankylosing spondylitis. *Analyst* 133(9), 1214–1220 (2008).

77 Adams JS, Hewison M. Unexpected actions of vitamin D: new perspectives on the regulation of innate and adaptive immunity. *Nat. Clin. Pract. Endocrinol. Metab.* 4(2), 80–90 (2008).

78 Julià A, Tortosa R, Hernanz JM *et al.* Risk variants for psoriasis vulgaris in a large case–control collection and association with clinical subphenotypes. *Hum. Mol. Genet.* 21(20), 4549–4557 (2012).

79 Arden N, Nevitt MC. Osteoarthritis: epidemiology. *Best Pract. Res. Clin. Rheumatol.* 20(1), 3–25 (2006).

80 Lamers RJ, Degroot J, Spies-Faber EJ *et al.* Identification of disease- and nutrient-related metabolic fingerprints in osteoarthritic guinea pigs. *J. Nutr.* 133(6), 1776–1780 (2003).

81 Malone DG, Irani AM, Schwartz LB, Barrett KE, Metcalfe DD. Mast cell numbers and histamine levels in synovial fluids from patients with diverse arthritides. *Arthr. Rheum.* 29(8), 956–963 (1986).

82 Tetlow LC, Woolley DE. Histamine stimulates the proliferation of human articular chondrocytes *in vitro* and is expressed by chondrocytes in osteoarthritic cartilage. *Ann. Rheum. Dis.* 62(10), 991–994 (2003).

83 Zhai G, Wang-Sattler R, Hart DJ *et al.* Serum branched-chain amino acid to histidine ratio: a novel metabolomic biomarker of knee osteoarthritis. *Ann. Rheum. Dis.* 69(6), 1227–1231 (2010).

•• **Uses a well-structured study design including a discovery phase and an independent validation phase to validate the metabolite associations with osteoarthritis.**

84 Felig P, Marliss E, Cahill GF Jr. Plasma amino acid levels and insulin secretion in obesity. *N. Engl. J. Med.* 281(15), 811–816 (1969).

85 Caballero B, Gleason RE, Wurtman RJ. Plasma amino acid concentrations in healthy elderly men and women. *Am. J. Clin. Nutr.* 53(5), 1249–1252 (1991).

86 Adams Jr SB, Setton LA, Kensicki E, Bolognesi MP, Toth AP, Nettles DL. Global metabolic profiling of human osteoarthritic synovium. *Osteoarthr. Cartilage* 20(1), 64–67 (2012).

87 Doherty M. New insights into the epidemiology of gout. *Rheumatology* 48(Suppl. 2), ii2–ii8 (2009).

88 Liu Y, Sun X, Di D, Quan J, Zhang J, Yang X. A metabolic profiling analysis of symptomatic gout in human serum and urine using high performance liquid chromatography-diode array detector technique. *Clin. Chim. Acta* 412(23–24), 2132–2140 (2011).

89 Cui Q, Lewis IA, Hegeman AD *et al.* Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotech.* 26(2), 162–164 (2008).

90 Tautenhahn R, Cho K, Uritboonthai W, Zhu Z, Patti GJ, Siuzdak G. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat. Biotech.* 30(9), 826–828 (2012).